

Dangers of going underground

Precautions need to be taken to protect researchers threatened by violent pressure groups. But open access to information and public accountability need to be sustained in the face of determined anti-rationality.

Faced with continuing attacks and threats from radical environmental and animal-rights groups, scientists increasingly find themselves lumped in with an odd assortment of societal targets — from fast-food restaurants to mink ranchers to second-hand car dealers — branded as agents of global capitalism and eco-destruction. No amount of academy studies proclaiming the safety of genetically modified (GM) crops, or protocols to guarantee the humane treatment of lab animals, or public-relations campaigns explaining the benefits of biotechnology to a hungry world, are likely to persuade the vandals and arsonists otherwise. They have reduced the world's complexity to a few simple manifestos, and have traded debate for intimidation.

So scientists should focus instead on winning over the vastly larger number of people who may be sceptical or uneasy about GM foods or the need for animal research, but who still have open minds. And the best way to do that is to stick with two of science's traditional strengths — the open exchange of information and a faith in rationality.

Two stories in this week's News section illustrate how difficult that high-minded road can be. A local planning board has rejected a proposal by the University of Cambridge to build a new primate research facility for fear of protests from animal-rights groups (page 725). And scientific societies are pressing the US agriculture department to remove information from the Internet about scientists who use laboratory animals — information that had been posted in the interests of transparency and public accountability (page 723).

The first action is the more lamentable, in that it represents a citizens' council simply caving in to bullying tactics. The second action, proposed by the research community itself, is defensible in spirit, but sets a worrying precedent. No one can blame scientists for wanting to protect themselves. The tactics of groups such as the Animal Liberation Front (ALF) and the Earth Liberation Front (ELF) have moved

well beyond spray-painting slogans and trampling crops, to arson. If that trend continues, it's only a question of time before someone gets hurt. So the request from the Federation of American Societies for Experimental Biology (FASEB) to remove detailed identifying information about animal research facilities from the web seems justified. And there should be a way to accomplish that while still satisfying the public's right to know. The US agriculture department, for example, posts information about GM crop trials on the web, as does the Department for Environment, Food and Rural Affairs in the United Kingdom. In neither case is the precise location of research plots given, presumably to make things difficult for would-be saboteurs.

Some 'secrecy', if it can be called that, therefore seems warranted. At the same time, scientists must scrupulously avoid giving even the slightest impression of taking their research underground. That would only further inflame the zealots, and make it harder to win over the uncommitted. Besides, it's not clear that arsonists rely that much on the kind of information posted on the agriculture department website. Groups such as the ELF or ALF don't seem especially diligent about doing careful research before they choose their targets. More likely, a short news story about a genetics lab being built at a nearby university, or the simple knowledge that a scientist uses animals — any animals — in their work, will suffice.

Take the case of Toby Bradshaw, the University of Washington plant researcher whose office was burned down last May by the ELF. The arsonists obviously misunderstood his work. Bradshaw has studied traditional hybridization of poplar trees, but has never done any genetic 'engineering'. And when the eco-vandals fire-bombed his office, they also destroyed 100 showy stickweed plants — a quarter of the remaining samples of one of the region's most endangered species. Which goes to show that if too much information is sometimes dangerous, too little can be even worse. ■

A boost too far?

Care will be needed over the proposed expansion of US bioterrorism research, lest the cure prove worse than the disease.

It should come as no surprise that President George W. Bush's "war budget" contains large new spending on research related to bioterrorism. The US government found itself in an untenable position after last autumn's anthrax attacks, unable to present a convincing case to the public that it knew what was happening, or how to respond. It couldn't tell postal workers what to do to protect their health or offer much advice on who else was at risk. If the attack had been sustained, it could have triggered mass panic in Washington and New York.

A government's first duty is to protect its citizens, and in those circumstances the United States has no choice but to invest heavily in bioterrorism research — if only to protect itself from future criticism. But this investment carries risks. As several researchers have already commented (see page 719), the rapid expansion of this activity will result in the circulation of more potentially dangerous materials and information to more people, even as the government attempts to

restrict the flow of both. It has to be asked whether any new regulatory regime can be smart and flexible enough to contain this contradiction.

Additionally, questions will surely be asked about the ability of the National Institute of Allergy and Infectious Diseases (NIAID) to absorb a massive increase in its budget without wasting money on second-rate science. It will take more than the suave assurances of Tony Fauci, NIAID's director, to make a convincing case that such an expansion can be managed efficiently.

It now falls to Congress to consider the proposal and to craft the final budget numbers. For the reasons noted above, no one there is likely to oppose Bush's bioterrorism spending plans outright. But it is Congress's duty to scrutinize the budget carefully, even when the nation considers itself to be at war. It would be appropriate if somebody there has the guts to ask some tough questions about the national-security implications of the proposed expansion. ■





Lost letters
Light shed on Bohr and Heisenberg's Copenhagen meeting
p720



Facing the facts
Unreviewed results creep into US cloning debate
p722



Switched off
Partner nations angered as US budget cut halts experiment
p723



Branching out
First tree genome will be sequenced by 2003
p725

Biodefence boost leaves experts worried over laboratory safety

Jonathan Knight, San Francisco

A massive infusion of cash for biodefence research, proposed with much fanfare in President George W. Bush's 2003 budget request on 4 February, has a number of US scientists and bioweapons experts on edge.

They worry that the new spending, which includes a sevenfold increase in biodefence funding for the National Institutes of Health (NIH), will stimulate the proliferation of labs that handle dangerous pathogens, and raise the risk of an accidental or deliberate release.

Nearly all of the NIH money for biodefence is to be administered by the National Institute of Allergy and Infectious Diseases (NIAID), whose budget would grow next year by almost 60%, to \$4 billion, under the Bush proposal. It includes \$430 million in new funds to build secure containment facilities for hazardous microbes, and \$533 million for drug and vaccine research. The NIAID would also get \$441 million, a sixfold increase, for basic research into the biology of bioterror agents.

But the prospect of dozens of labs across the country handling pathogens such as those that cause anthrax and tularemia is nothing short of terrifying, says Richard Ebright, a molecular biologist at Rutgers University in Piscataway, New Jersey.



The bigger picture: President Bush's crusade against bioterrorism has provoked concern among biodefence experts that the chances of an accidental or deliberate pathogen release may be increased.

Ebright says he developed an intense interest in biodefence issues after the anthrax attacks on the United States last October were launched from his state. He argues that although detection techniques are critical for good biodefence, understanding the basic

biology of the pathogens is not. "From a security perspective, we would do better to have much more restricted access and less information," he says.

Officials at the NIAID have defended the increase from criticisms that it might cause

American Red Cross turns its back on stem-cell grant

Jonathan Knight, San Francisco

The American Red Cross has refused to accept the first federal grant to be awarded in the United States for research using human embryonic stem cells.

The award was a \$50,000 supplement to an existing five-year grant for leukaemia research using mouse embryonic stem cells, given to a team led by Robert Hawley of the Red Cross's Holland Laboratories in Rockville, Maryland.

After the National Institutes of Health announced the award on 8 February, American Red Cross chief scientific officer Jerry Squires issued a statement saying that

the organization had revised its research priorities and would decline the money.

This explanation met with deep scepticism from scientists familiar with using stem cells. To refuse a grant already awarded is extremely rare, says Lawrence Goldstein, a cell biologist at the University of California, San Diego, who uses mouse embryonic stem cells. "Why, after going through all the trouble to apply, they would turn around and refuse the money is beyond me," he says.

Hawley's five-year grant, which lasts until 2006, is to study genes involved in the differentiation of mouse blood precursor

cells. According to Goldstein, confirming the results in human cells would be a logical extension of such work. Hawley did not respond to a request for an interview.

Several researchers speculated that the American Red Cross — still smarting from a public row over its distribution of funds gathered after the 11 September tragedy, and the subsequent departure of its president, Bernadine Healy — wanted to avoid being the first to take funds for research using human embryonic stem cells. "Maybe they just didn't want one more controversy," says Sean Tipton, a spokesman for the American Society for Reproductive Medicine.

security concerns. "The infrastructure that will be put in place will make it extraordinarily difficult for there to be a deliberate release of biological pathogens," says NIAID director Anthony Fauci.

Nancy Connell, director of the Center for Biodefense at the New Jersey Medical School in Newark, says there is no justification for such a massive increase in spending. "Those kinds of big numbers mean going deeper into the competition," she says. "We will be starting to fund some bad science." She also questions whether adequate security can be maintained if the number of labs handling bioterror agents shoots up.

But security may not be of equal concern for all hazardous biological agents, argues Steven Block, a biophysicist at Stanford University in California and a member of the JASON group, which advises the government on technical issues related to defence. Whereas the smallpox virus may be hard to come by, he points out, deadly microbes such as *Escherichia coli* O157:H7, which can contaminate food, the bacterium *Francisella tularensis*, responsible for tularaemia, which flared up two years ago in Kosovo, and even the bacterium that causes anthrax, *Bacillus anthracis*, can be readily obtained from nature.

This distinction may soon be enshrined in law. A bill sponsored by Senators Edward Kennedy (Democrat, Massachusetts) and Bill Frist (Republican, Tennessee) instructs the health department to work with scientists in drawing up a new list of biological agents that should be subject to tight regulation. Ron Atlas, president-elect of the American Society for Microbiology, says he would hope for a hierarchy of security requirements that would take the varying risk of different agents into account.

Block and others point out that most biological agents pose only a limited threat without the technology to turn them into weapons them. But research into this would also expand under the Bush budget plan, which includes \$600 million in new funding related to bioterrorism at the Department of Defense. Part of this would fund studies of "how potential bioterrorism pathogens may be weaponized, transported, and disseminated"; according to a budget fact sheet distributed by the White House.

From an international perspective, the Bush biodefence plan is part of a worldwide trend towards greater funding of biodefence research, says Jean Pascal Zanders of the Stockholm International Peace Research Institute. At least 12 countries, including Sweden, already have high-containment facilities, known as biosafety level 4 labs, and several others are expected to announce plans this year to build their own, Zanders says. "We could get a sort of arms race on the defensive side," he says, "and it might spill over into an offensive one."



Happier times: Bohr (right), who regarded Heisenberg almost as a son, later wrote him angry letters.

Physicist's letters reveal clues to bitter wartime rift

Alison Abbott, Munich

Unsent letters written by Danish physicist Niels Bohr to his German protégé Werner Heisenberg saw the light of day for the first time on 6 February. But the long-awaited publication has failed to solve the mystery of what happened during their unhappy meeting in Nazi-occupied Copenhagen.

Heisenberg, who headed Germany's unsuccessful attempt to produce a nuclear bomb, visited Bohr in September 1941 with a colleague, Carl Friedrich von Weizsäcker, on a cultural-exchange mission to Copenhagen.

Heisenberg and Bohr went for a walk together, during which Heisenberg brought up the subject of the German nuclear-bomb programme. There is no consensus as to what exactly was said, or what Heisenberg's motive for the visit had been — whether he was trying to recruit Bohr to the German programme, trying to find out how far the Allies had progressed in their attempts to build an atomic bomb, or taking the first steps towards a moratorium against nuclear weapons. All that is known is that the incident caused a deep rift between them. The mystery formed the basis of Michael Frayn's successful play *Copenhagen*.

The newly published letters and notes were written between 1957 and 1962. After the war, Heisenberg claimed that he had wanted to discuss with Bohr the prospect of a worldwide moratorium on bomb development programmes, and moreover that he had deceived the Nazi regime into believing that a bomb was not technically possible.

But in one letter, Bohr says he remembered Heisenberg telling him "everything was being done in Germany to develop atomic weapons" and that Heisenberg "had spent the last two years working more or less exclusively on such preparations". In another, Bohr writes: "It is therefore quite incomprehensible to me that you should think that you hinted to me that German

physicists would do all they could to prevent such an application of atomic science."

Immediately after reading the letters, von Weizsäcker, whose brother Richard was the German president in 1984–94, told the German Press Agency that Bohr's memories "contained deep errors". In an interview with the Munich newspaper *Süddeutsche Zeitung*, he said that Heisenberg had indeed been on a peace mission to promote a bomb moratorium, albeit prompted partly by self-interest. Because Germany had given up its own programme, Heisenberg and von Weizsäcker wanted the Allies to do the same, "so that a bomb wouldn't fall on us".

Very few of the facts and viewpoints put forward are totally new to historians, but Finn Aaserud, director of the Niels Bohr Archive in Copenhagen, points out that they indicate the depth of agitation of the scientists involved. "Bohr clearly worried very intensely about it — he had, after all, considered Heisenberg as a sort of son." Thomas Powers, who wrote *Heisenberg's War: The Secret History of the German Bomb*, the book on which *Copenhagen* was based, agrees. "For the first time we have a clearer insight into what made Bohr so angry," he says.

Aaserud also notes that Bohr's letters were written during the cold war, many years after the meeting, and that seeds of doubt about Heisenberg's motivation could have been planted by the questions asked him by intelligence services and historians.

Helmut Rechenberg, head of the Heisenberg Archive in Munich, and one of Heisenberg's past pupils, comments that the letters show how much Bohr wanted to discuss with Heisenberg the meeting that disrupted their relationship. "And that's a pity, because Heisenberg told me, and others, that he wanted very much to discuss it with Bohr," says Rechenberg. Bohr's death in 1962 made this impossible.

♦ www.nbi.dk/NBA/papers/introduction.htm

Rockefeller head quits as scandal looms

Erika Check, Washington

Arnold Levine has resigned as president of Rockefeller University in New York amid reports of a public and allegedly inappropriate encounter with a female graduate student.

In a statement issued on 10 February, 62-year-old Levine said he was leaving to address matters affecting his health. "In light of my health issues, I regret I will not be able to continue to lead this extraordinary institution and these talented people," he wrote.

But sources at the university confirmed reports that Levine has resigned over what one of them described as an "inappropriate incident" with a female graduate student that occurred on 10 January. Various reports of the encounter have swept through Rockefeller in recent weeks.

According to the sources, the encounter occurred in the university's faculty club, a popular meeting place for researchers and graduate students. Both Levine and the participating student were intoxicated, the sources say, and she is said to have told university officials that the encounter was consensual. A male student who was in the faculty club confronted Levine during the encounter, one of the sources says, at which point Levine became angry. This student reported the incident to the Board of Trustees.

Levine, who was the first to isolate the p53 tumour-suppressor protein, rushed back on 8 February from a meeting held to mark the opening of the Spanish National Cancer

Centre in Madrid. Summoned by the Board of Trustees, he expressed profound regrets and said he would resign, says a source familiar with the situation. Levine was unavailable to comment on the incident.

Thomas Sakmar, who works on mechanisms of cell signalling and has headed the university's Academic Senate for the past year, has since been appointed acting president while the board sets about selecting a permanent replacement for Levine.



Arnold Levine resigned as president soon after a meeting with Rockefeller's trustees.

In a statement, Richard Fisher, chairman of Rockefeller's Board of Trustees, called Levine an "admirable and inspirational leader" and said Levine had "helped strengthen the university and position it for continued pre-eminence as it enters its second century".

The resignation has shocked and saddened many inside and outside Rockefeller's campus on Manhattan's upper east side.

Harold Varmus, former director of the National Institutes of Health and current president of the Memorial Sloan-Kettering Cancer Center in New York, says that Levine had been instrumental in strengthening the research community in New York. "There will be an empty ache not having Arnie at the helm across the street," Varmus says.

Levine left Princeton University in 1998 to lead Rockefeller. He is widely credited with reinvigorating the small but prestigious university, which had suffered a blow to its morale with the forced resignation of David Baltimore as president in 1991. Baltimore left amid charges that a co-author had falsified data in a 1986 paper — a charge that was dismissed by a federal appeals panel in 1996. He is now president of the California Institute of Technology in Pasadena.

Baltimore says that Levine was successful at Rockefeller, recruiting good people and raising money effectively. "I think Arnie's been doing a terrific job," Baltimore says. "Rockefeller is the kind of institute he understands, and he'll be a very hard person to replace." ■

S. CHERNIN/AP

Soros offers open access to science papers



George Soros wants free information.

Declan Butler

George Soros, the Hungarian-born financier and philanthropist, is backing a new effort to provide free and unrestricted access to scientific and other academic literature.

Soros's Open Society Institute (OSI) will launch the new initiative in Budapest on 14 February, pledging US\$3 million in grants over three years to support free electronic article repositories and 'alternative' journals committed to open access.

Peter Suber, a philosopher at Earlham College in Richmond, Indiana, and one of the architects of the 'Budapest Open Archive Initiative', says that this sum will go a long way. He notes that electronic repositories are

not expensive to set up, and that the open-source software required to run such repositories is already freely available on the Internet.

But Suber and others involved in the initiative concede that further funding will ultimately be required, and that their broader aim is to create a "domino effect" by convincing other funders and research institutions to become involved.

The initiative is publicly supported by around 300 individuals, including the provosts of the California Institute of Technology and the University of Kansas. The 20 organizations listed as supporters include the Public Library of Science, the University of Illinois at Urbana-Champaign and the University of Missouri-Columbia.

The initiative shares some of its objectives with the Public Library of Science, which last year sought to organize a researchers' boycott of journals that refused to cooperate with its campaign for free

access to scientific literature. But instead of seeking to influence publishers, the Budapest initiative will attempt to win support for open-access publishing from within the academic community.

Research institutions and funding agencies that sign up commit themselves to making policy changes, such as creating local open-access electronic repositories, and making it compulsory for grant recipients to deposit their papers there. Individual signatories agree to deposit their research in freely available electronic repositories, and to support alternative journals as authors, editors and referees.

"Only time will tell if the impact of the OSI will be large or small," says Suber. "We only need enough money to change the business model of enough scholarly resources to give momentum to this better way. The money already in the system is more than enough to pay for open access," he adds. ■

www.osi.hu

S. GALLUP/GETTY

Cloning agenda 'skewed' by media frenzy

Erika Check, Washington

As the debate over human cloning in the United States gathers pace, scientists are becoming increasingly concerned that research that has not been peer-reviewed is exerting undue influence on the country's legislators.

Their concern was emphasized at a Senate judiciary committee hearing on cloning held on 5 February. The discussion centred on therapeutic cloning — whether cells from patients should be used to create cloned embryos as a source of stem cells, which could then potentially be grown into replacement tissues perfectly matched to the patient.

At the hearing, Senator Sam Brownback (Republican, Kansas), author of a bill that would ban cloning for any purpose, said that such cloning is unnecessary. His argument was based, in part, on reports that Catherine Verfaillie, a developmental biologist at the University of Minnesota, has grown functional differentiated cells from adult stem cells.

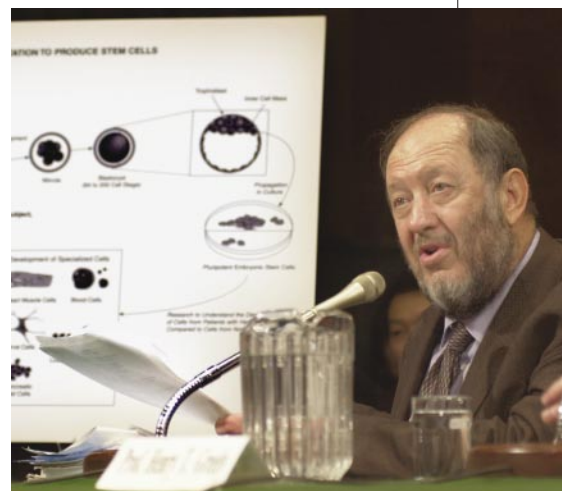
At the same hearing, Brownback's opponent on the issue, Dianne Feinstein (Demo-

crat, California), cited a claim that scientists had constructed a kidney from a cloned cow embryo as evidence of the promise of therapeutic cloning.

But neither piece of work has been reviewed or reproduced by other scientists. Such unreviewed claims are cropping up with increasing regularity in policy debates, thanks in part to the rapid dissemination of information on the Internet. This raises the question, researchers say, of whether policy-makers know — or want to know — the difference between a claimed result and a peer-reviewed scientific finding.

Verfaillie's work on adult stem cells was described in *New Scientist* on 26 January. The magazine reported the story on the basis of a patent related to her work.

The cloned-kidney claim first surfaced on 4 December in the *Los Angeles Times*, when scientists at Advanced Cell Technology (ACT) in Worcester, Massachusetts, told a journalist about the supposed advance. In late January, the claim resurfaced in British and American newspapers.



Printed problems: Irv Weissman says that the media 'crossed a line' in the cloning debate.

K. CEDENO/AP

Stanford biologist Irv Weissman urged senators at last week's hearing to disregard these reports. Indeed, he said, Verfaillie has stated in a letter she sent to legislators that her research should not be used to justify a ban on therapeutic cloning. Researchers will "not know which stem cells, adult or embryonic, are most useful in treating a particular disease without side-by-side comparison of adult and embryonic stem cells," she says in the letter, adding that she supports research into therapeutic cloning.

Although scientists have not questioned the accuracy of Verfaillie's work, they say it would be premature to cease work on embryonic stem cells because it is unclear whether cell lines made from adult stem cells will prove as long-lived or function as well as cell lines made from embryonic tissue.

The ACT work, meanwhile, is of dubious relevance to therapeutic cloning because the cells used to grow the kidney came from an embryo that had been allowed to develop to an early fetal stage. This would be ethically unacceptable if human embryos were used.

Michael West, chief executive of ACT, has apologized for his company's leak of the cow kidney data. "It wasn't intentional," he says.

Nevertheless, scientists are worried that these leaks are unduly influencing the legislative debate. "The fear is that policies will be made based on premature publicity," says Helen Blau, director of the Baxter Laboratory for Genetic Pharmacology at Stanford University.

Researchers have roundly denounced the media for printing unreviewed claims. Weissman says the media "has crossed a line" by printing such reports. But other researchers concede that scientists have a responsibility to refrain from disseminating or commenting on work that has not been peer-reviewed.

Share crash puts focus on accounts

Erika Check, Washington

The plummeting share price of a leading Irish drug company could lead to closer scrutiny of how such firms account for their research and development spending, analysts say.

Elan, based in Dublin, was until recently the largest company on the Irish stock market, and is also listed on the New York Stock Exchange. Its stock price, which fell from a high of \$63 last July to \$44 at the beginning of the year, has dropped by 70% since then, driven by a number of factors — including comparisons of its accounting practices with those of Enron, the failed US energy trading company.

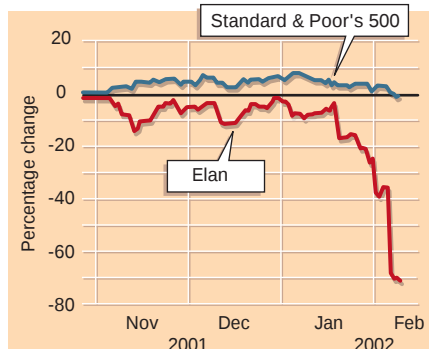
Elan may have stumbled, but it is not about to fall: it reported sales of \$350 million

last year, and has several products in late-stage clinical trials. "We've got \$2 billion in the bank and another \$2.4 billion in marketable securities," says a company spokesman.

But analysts say that Elan's example is a warning to other drug companies that use some of its accounting techniques. According to an investigation published in the *Wall Street Journal* on 30 January, Elan had set up 55 joint ventures, structured so that it could charge research costs to its partners, instead of counting them as part of its own expenses. It also charged the partners a licensing fee, which Elan counted as income. To wary investors, this set-up looks not unlike that used by Enron to hide its debt.

Analysts say that many drug and biotechnology companies are involved in joint ventures. "If Elan had had five of these it would have been okay," says Ian Sanderson, an analyst at SG Cowen Securities in Boston, "but it had 55 of them generating total revenues of about \$200 million a year."

Elan's announcement on 18 January that four French patients had become ill during a clinical trial of its drug for Alzheimer's disease (see *Nature* 415, 462; 2002) initially pushed its stock price down (see chart). This was compounded by disappointing financial results released on 4 February, which included new details about Elan's accounting methods — triggering an investigation by the US Securities and Exchange Commission.



Falling stock: Elan's share price has plunged against the US average since mid-January.

Anger at US plan to drop physics experiment

Geoff Brumfiel, Washington

Plans to shut down an international high-energy physics experiment have prompted an angry response from the project's foreign partners, some of whom say the decision could endanger future international collaborations with the United States.

The experiment at Brookhaven National Laboratory in New York state, known as E949, is designed to spot very rare types of particle decay. These can be used to measure the weak force that acts between quarks. Only two such decays have been observed, in other experiments, over the past decade. E949 was expected to capture up to 20 more decays in the next two years, but is now likely to end in April.

The Department of Energy (DOE), which funds the Brookhaven laboratory, announced plans to end the project last week after receiving no increase in its budget for the 2003 financial year (see *Nature* **415**, 564; 2002). The money saved will help to finance upgrades to the giant Tevatron accelerator at Fermilab, near Chicago.

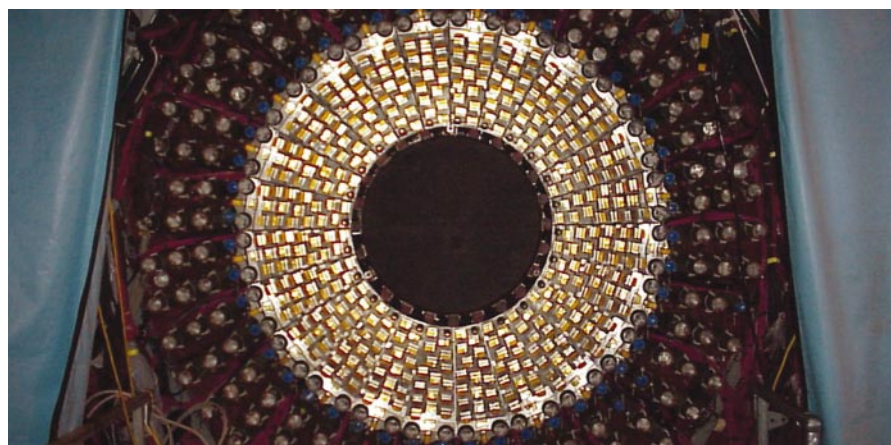
The two-thirds of E949's researchers who came from outside the United States seem to have been left in the lurch. "Our group has built tracking devices, high-speed electronics and many other detectors, all in the expectation that this would be carried out as planned," says Doug Bryman, a physicist at the University of British Columbia in Canada and spokesperson for the project.

Shojiro Sugimoto, from the KEK accelerator in Tsukuba, Japan, adds that several PhD students from Tokyo and Kyoto universities may have to abandon their theses.

Jean-Michel Poutissou, associate director of TRIUMF, Canada's national laboratory for particle and nuclear physics, says that the decision could affect future projects. "I think it will have a major impact on the way we deal with American collaborations," he says. Poutissou points out that E949's Canadian contingent will soon go before a government committee to request funding for the next three years. "How can they make a decision if we don't know the host laboratory will make good on the promises made?" he asks.

US officials acknowledge that the decision had caused difficulties for their partner nations. "It is regrettable," says Jim Decker, the acting director of the Office of Science at the DOE. "We particularly don't like to do this in a situation where researchers from other countries are affected, but sometimes it becomes unavoidable."

But a senior official at the National Science Foundation says he doubts that the decision will damage future international projects, such as the proposed Next Linear Collider. "I don't think it will have a direct impact on other collaborations," he says. "There's a long history of experiments ending before their time."



Brookhaven blues: the E949 experiment to detect rare particle decays could end within months.

L. LITTENBERG/BNL

Researchers fear web information will prompt attacks

Tony Reichhardt, Washington

Scientific societies are urging the US Department of Agriculture to give less information on its website about researchers who use laboratory animals.

Robert Rich, president of the Federation of American Societies for Experimental Biology (FASEB), wrote to the department's Animal and Plant Health Inspection Service (APHIS) on 6 February, asking it to revise its procedures on reports by animal-welfare inspectors. These have been posted on the Internet since last October, in accordance with the 1996 Electronic Freedom of Information Act.

Rich warned that some reports include "excessive detail" about researchers, including the addresses of their laboratories, that "may expose individuals to harassment from animal-rights terrorists". He also expressed concern that some information might be inaccurate or preliminary.

Several other groups have joined FASEB in requesting that information allowing the



identification of individuals or laboratory locations should not be released.

George Nethercutt (Republican, Washington), who sits on the House of Representatives agriculture appropriations subcommittee and science committee, has asked APHIS "to work closely with the research community to resolve this matter".

His spokesman, Robert Neal, says that there could be "legislative remedies" if the service does not act voluntarily. "We don't have to make it easy for people who are clearly hostile to research," he says.

Nethercutt was scheduled to testify this week at a congressional hearing on 'ecoterrorism' by the forestry subcommittee of the House Resources Committee, chaired by Scott McInnis (Republican, Colorado).

Groups such as the Earth Liberation Front (ELF) — which claims to have inflicted \$30 million in property damage against companies and universities it accuses of harming the environment — regularly target biotechnologists, according to Neal. The ELF has claimed responsibility for arson (see *Nature* **411**, 509; 2001) and other attacks on biotech facilities.

Nethercutt has proposed legislation that would stiffen the criminal penalties for such activities. It would also give money to the National Science Foundation to study improved security measures.

Satellite orbit failure DASHes hopes for Japanese launcher

Tokyo Japan's space programme suffered another setback last week when a satellite failed to separate from its launch rocket.

The DASH satellite had been scheduled to orbit Earth for three days, and to spend its return journey collecting data on the high temperatures experienced during re-entry. The ¥600-million (US\$4.5-million) spacecraft failed to respond to signals after lift-off on 4 February and has now been written off. Another satellite was successfully placed in orbit using the same H-IIA rocket.

Investigations are now under way to find out whether the satellite or the launch rocket was to blame for the failure. The problem is the latest to afflict the H-II programme, which is attempting to prove its reliability as a launcher of commercial satellites.

► www.nasda.go.jp/index_e.html

Russian spy trial put on hold in hunt for evidence

Moscow The espionage trial of a Russian physicist was adjourned last week so that prosecutors would have more time for their investigations.

Valentin Danilov, head of the thermophysics centre at Krasnoyarsk State Technical University in Siberia, is accused of selling technical information about satellites to a Chinese company. The Russian Federal Security Service claims that the material was classified and could be used to develop space weapons. If convicted, he faces up to 20 years in prison.

Danilov has argued that the material was declassified in 1992 and published in Russian scientific journals. Colleagues in the Russian Academy of Sciences have called for a public trial and an independent investigation of the charges, which they say are absurd.

Molecular biology mourns leading light

London Max Perutz, the Austrian-born Nobel laureate who unravelled the structure of haemoglobin, has died at the age of 87.

A key player in the scientific revolution of the 1950s and 1960s that established the field of molecular biology, Perutz showed that the functions of proteins can be understood by elucidating their structures. He shared the 1962 Nobel Prize in Chemistry with Sir John Cowdery Kendrew, a colleague at the Laboratory of Molecular Biology at the University of Cambridge.

Perutz was renowned for his enormous



Protein pioneer: Max Perutz has died, aged 87.

capacity for work at the bench. Molecular biologist Jacques Monod, a fellow Nobel laureate, once commented that Perutz was one of the few scientists whose work grew more interesting as he got older. Perutz published his last paper in July last year (*Nature* **412**, 143–144; 2001).

Targeted funding angers Australian academics

Sydney One-third of the Australian Research Council's grant for the next five years is to be channelled into four specific research areas, the country's federal government announced late last month. About A\$170 million (US\$87 million) will go to projects in materials science, complex systems, genomics and gene expression, and photonics.

The decision, which comes just six weeks before the council's deadline for applications for 2003 funding, has angered some academic

groups, who say that they were not consulted. "There is absolutely no publicly available information for the rationale of these choices," says Michael Barber, secretary for science policy at the Australian Academy of Science and pro vice-chancellor of the University of Western Australia in Perth.

Vicki Sara, chief executive officer of the council, says the decision to prioritize certain projects is in line with the government's innovation plan, announced last year.

► www.arc.gov.au

Primate centre plans halted by protest fears

London Fears of protests by animal-rights activists have derailed plans for a new primate-research facility at the University of Cambridge.

The local council rejected a planning application for the centre on 6 February after the police warned that the laboratory would lead to "demonstrations that will result in road blockages and a serious danger to public safety". Regular protests have been mounted outside the Cambridgeshire drug-testing company Huntingdon Life Sciences.

Barbara Davies, a spokesperson for the Research Defence Society, which defends the use of animals in research, says the decision sets a worrying precedent. "It sends a message

to these minority groups that protests can have an effect and influence decisions," she says. The university says it may appeal. The application was refused once before, in January last year, on environmental grounds.

Geneticists branch out to sequence poplar genome

Washington The first tree genome should be sequenced by 2003, say officials at the US Department of Energy.

Plant geneticists say that the tree, a member of the genus *Populus*, which includes poplars and aspens, was chosen because it is heavily studied and has a relatively small genome.

Toby Bradshaw, a molecular geneticist at

the University of Washington in Seattle, says that researchers could use the *Populus* genome to learn how woody plants evolved. It may also help them to engineer faster-growing trees, or trees that can aid the decontamination of hazardous waste sites.



Poplar choice: genome sequence by 2003.

The sequencing work will be conducted by scientists at the Department of Energy's Joint Genome Institute in Walnut Creek, California, Oak Ridge National Laboratory in Tennessee, the University of Washington, and facilities in Canada and Sweden.

Bottom's up for the Bard's love potion

London The legendary aphrodisiac from William Shakespeare's *A Midsummer Night's Dream* has been recreated to celebrate St Valentine's Day. The potion was prepared by researchers at the fragrance company Quest International, based in Naarden in the Netherlands, following a request from Britain's Royal Shakespeare Company to the other RSC, the Royal Society of Chemistry.

The elixir — which promises to "make man or woman madly dote upon the next live creature that it sees" — was prepared from the flower *Viola tricolor*, a member of the pansy family. "It will be interesting to see what kind of effect the perfume might have," says Charles Sell, an organic chemist with Quest. The potion certainly gave Titania, the queen of the fairies and its first recipient, an interesting time: she fell in love with Bottom, whom previous magical shenanigans had left with the head of a donkey.

DOE



The counting house

Scientists' work is often evaluated using citation statistics compiled by a company called the ISI. But how useful and reliable are the data? David Adam gets the measure of citation analysis.

There are, it is said, three types of lies: lies, damned lies and statistics. Many scientists, who find their work assessed through attempts to gauge how often it is cited in the scientific literature, would surely subscribe to that view. Citation analysis, in the hands of non-experts, can be an extremely blunt instrument. What's more, the specialists in the field have found that raw citation data often contain errors.

Practitioners of citation analysis rely on data gathered by the ISI, a company in Philadelphia formerly known as the Institute for Scientific Information. Over the past four decades, the ISI has scanned reference lists in scholarly publications and collated citations to previously published work. The resulting database was actually developed for information retrieval — allowing

researchers to conduct rapid literature searches and to identify individual scientists working on particular topics.

Because it is hard for governments, funding agencies and promotions committees to find reliable yardsticks for measuring research quality, however, they often use the ISI's citation data to help them perform such evaluations. Important papers, the argument goes, will be cited more frequently. As a general rule, that is a reasonable assumption. But apply it blindly, without regard to the quality and limitations of the raw data, and the conclusions you draw may be far from reasonable.

"You must be very careful because it's about the reputation of scientists," says Anthony van Raan, director of the Centre for Science and Technology Studies at Leiden University in the Netherlands. "A database

prepared for information retrieval is not 100% suited to evaluation."

The ISI has always acknowledged the limits of citation analysis. But the lure of numbers has proved irresistible to those charged with judging scientists' work. And since the ISI was sold by its founder, Eugene Garfield, in the early 1990s, it has reacted to this demand. Now owned by the Thomson Corporation of Toronto, the company is producing software packages to help users probe its database, including one launched last year, called Essential Science Indicators, that promises to evaluate "potential employees, collaborators, reviewers, and peers".

At the same time, the ISI has toughened its attitude on the use of its data by independent bibliometrics researchers. As a result of Thomson's more hard-headed business

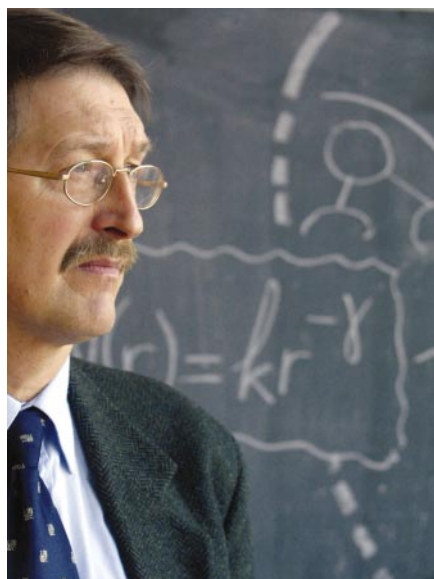
stance, some groups complain that they now face price rises and restrictions on data use that may force them out of citation analysis altogether.

Many bibliometrics researchers are concerned about this trend because they claim that the independent groups help to provide quality control for citation data. Others warn that the ISI is introducing products that are ripe for misuse. "They cover themselves by addressing all the issues in the accompanying notes, but who ever reads those?" asks one bibliometrics researcher. "It is frustrating for units such as ours."

Hit and miss

As an example of what can happen when citation statistics are misused, many experts point to the ISI's journal impact factors — a measure of the average number of citations garnered by the papers that each journal contains. Publishers have eagerly latched onto these data, using favourable impact factors in promotional material for their journals, and librarians have found them to be a convenient guide in deciding which journals to subscribe to.

But the use of journal impact factors has gone much further, extending to the evaluation of individual institutes, departments and scientists. The most obvious measure of the interest in various researchers' work would be to count citations to their papers directly. That is possible from the ISI's data, but getting the ISI or an independent group to conduct the analysis has historically proved expensive and time-consuming. So as a cheap-and-cheerful alternative, many evaluating bodies look at scientists' publication records and evaluate the quality of their output in terms of the impact factors of the journals in which their papers appear — figures that are readily available. It's "the poor man's citation analysis", says van Raan.



Anthony van Raan sees evaluations using journal impact factors as 'the poor man's citation analysis'.

Universities in Germany, for instance, regularly plug the impact factors of journals in which scientists publish into formulae to help them determine departmental funding. The Italian Association for Cancer Research requires grant applicants to complete worksheets calculating the average impact factor of the journals in which their publications appear. Elsewhere, the implicit use of journal impact factors by committees determining promotions and appointments is endemic.

Field trips

So why the controversy? Draw up a list of journals in a particular field and, with a few exceptions, there seems a pretty good correlation between a journal's impact factor and its perceived quality. But start making comparisons between fields — something the ISI warns against — and the results quickly become meaningless: mathematics researchers rarely cite more than one or two references, for example, whereas a typical paper in molecular biology includes dozens. This causes a wide variation in impact factors, even between comparable journals serving different disciplines, notes Per Seglen, a cancer researcher who also works on bibliometrics at the Norwegian Radium Hospital in Oslo, Norway (see chart, left). The figures are also biased in favour of journals that predominately publish review articles, which tend to be cited more frequently.

Other problems are less obvious, but when it comes to evaluating individual scientists, potentially more serious^{1,2}. Seglen points out that about 15% of the articles in a typical journal account for half of the citations gained by that publication (see chart, right). This means that a typical paper in a journal with a high impact factor may

not, in fact, be cited much more frequently than the average paper in a lower-ranking journal. "There is a general correlation between article citation counts and journal impact, but this is a one-way relationship," Seglen says. "The journal does not help the article; it is the other way round."

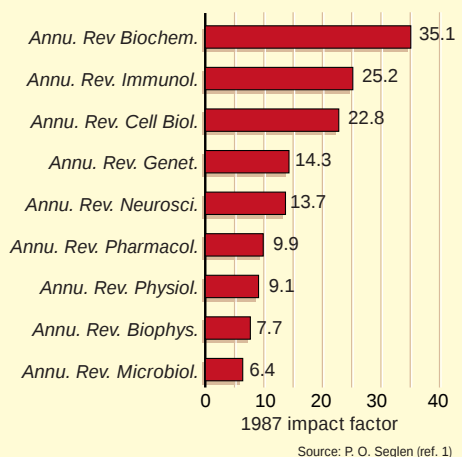
Again, the ISI cautions against using journal impact factors to evaluate individuals. "These scores were never designed by ISI to be proxies for the influence of papers or, when aggregated, the work of individuals," says David Carter, the company's vice-president for corporate communications.

But that has not halted the trend, of which Finland provides an extreme example. There, government funding for university hospitals is partly based on publication points, with a sliding scale corresponding to the impact factor of the journals in which researchers publish their work.

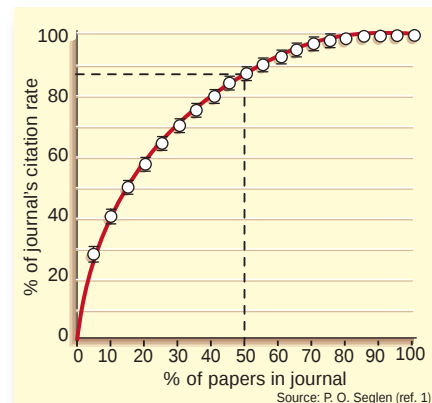
"To my knowledge, Finland is the only country in which the journal impact factor has been canonized in the law of the land," says Kari Raivio, rector of the University of Helsinki. Raivio calculates that a single paper published in a journal with an impact factor of 3, rather than 2, could have boosted a hospital's funding by about US\$7,000 in 2000.

Uncertain ratio

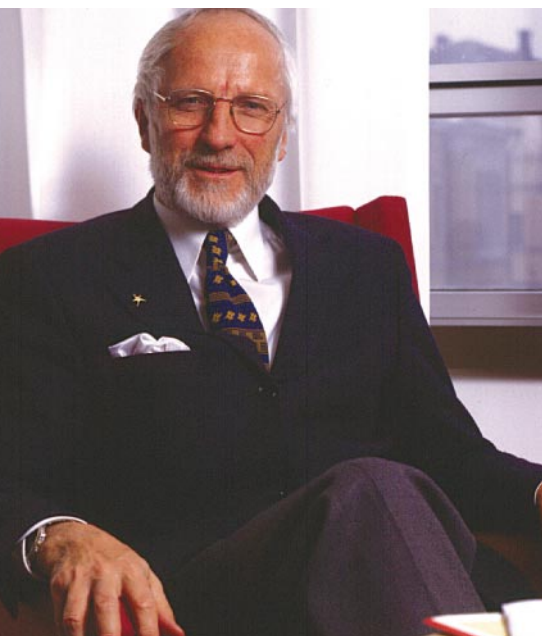
The ISI calculates a journal's impact factor for a given year by searching its database for the number of citations that year to articles published in the journal in the preceding two years, and dividing by the number of 'citable' papers published by the journal in those two years. But this can raise another problem because the numerator in the equation can include citations to articles that do not appear in the list of citable items — generally restricted to original research papers and review articles. According to Henk Moed, a researcher at van Raan's Leiden centre, this can produce an impact factor that is up to 40% too high in some



Different strokes: the impact factors vary between comparable journals in various disciplines.



Uneven split: about half the citable papers in a journal account for over 85% of its citations.



Finland is the only country in which the journal impact factor has been canonized in law. Kari Raivio

title, volume, year and page number. Next, a computer takes a few bytes of information from each highlighted field to build up an identifying code or 'tag' that is unique to that paper. A similar data-capture and tagging process occurs for the references at the end of the paper. Algorithms then compare the citation tags with any article tags already in the database, and each successful match counts as a citation.

Journal impact factors are calculated using a simpler method. Of the various fields highlighted in each reference, only the title of the journal and the year are used. This generates a bulk count of references to a particular journal in any given year, but means that citations cannot be matched directly to individual articles. This helps to boost the impact factors of journals — such as *Nature*, *Science* and several leading medical journals — that include many different sections, including

news and correspondence pages, which may gain citations but are not 'citable' papers.

These biases are systematic, but investigations by *Nature* suggest that the ISI's journal impact factors can, on occasion, be subject to less predictable problems, arising from fluctuations in the counts of 'citable' items. For *Nature Genetics*, there appears to have been a significant undercount of citable items in 1996. More recently, the count of citable items for *Nature* in 2000 seems to have been inflated by the erroneous inclusion of items other than original research reports and review articles — including some from a section called Futures, which featured short science-fiction stories.

Blazick admits that some mistakes are possible as the 'citable' and 'non-citable' articles must be separated manually. "There's some room for error with how fast we do the process," she says, claiming that errors are likely to even out, with as many citable articles missed as non-citable ones included.

The ISI declined to comment on specific cases, but its officials say that the company does address errors that are drawn to its attention and is always striving to improve accuracy. "We are constantly coordinating with our publishers to enhance our processes," says Carter.

One currently live issue is the development of procedures to collate citations to

PHOTO NERO

cases. To understand why, it is necessary to know how the ISI counts citations.

Pamela Blazick, editor of the ISI's *Journal Citation Reports*, in which journal impact factors appear, says that pages from journals are first scanned using optical character-recognition software. To store a research paper in its database, ISI employees highlight the following fields: author, address, journal

Pretenders to the throne

The ISI is the undisputed king of the counting houses, but new software that scans articles, pulls out references and automatically generates citation indexes may challenge its monopoly, as publishers and digital libraries start to adopt these tools.

Like the ISI's database, these new products have arisen from efforts to improve the accessibility of the academic literature. CrossRef, for instance, is a non-profit consortium — to which 99 publishers have signed up — that allows users to click straight from the references in online papers to the abstracts or full texts of some 3.9 million research articles in more than 5,600 journals.

To achieve such a set-up requires software, manual systems or a mixture of the two that can accurately identify all references — which will be written in a variety of formats and are often replete with mistakes. Such systems must also identify elements such as author names and starting page numbers. Each reference is then compared with those in a central

database to find an identifier — a URL for the web page of the paper concerned, or the paper's unique digital object identifier (DOI).

But once references and papers are interlinked, it is relatively simple to apply conventional algorithms to create citation indexes, impact factors and other indices, and to rank cited researchers and articles, much as the ISI has been doing for years.

CrossRef has so far made no moves in this direction — but other groups have. Among the pioneers are the developers of ResearchIndex at the NEC Research Institute in Princeton, New Jersey. This tool has been used to build, without human intervention, a digital library of more than 300,000 papers on computer science gathered from the web. ResearchIndex is not error-free — in tests, it has matched references to articles with about 95% accuracy; for journals with standard citation practices, its performance should be close to 100%. But project leader Steve Lawrence argues that the system could be expanded to cover a wider range of publications than the

ISI's database, which is more cost-intensive because of its manual inputs.

Fully automated citation-indexing systems are proliferating. The Open Citation Project is building one for electronic preprint archives, and NASA's Astrophysics Data System (ADS) has one for its collection of 3 million abstracts. Günther Eichhorn of the Smithsonian Astrophysical Observatory in Cambridge, Massachusetts, project scientist for the ADS, says that users are increasingly using the system to create their own citation statistics.

To improve results, companies such as Parity Computing of San Diego are using a hybrid approach, automating most functions, but also employing people at key points to increase quality control. Chris Rosin, Parity's president, claims that its system achieves over 99% accuracy. The company recently implemented a citation index for the 50-year literature archive of the Association for Computing Machinery, the world's largest computing society. Parity intends to build on such citation

indexes to generate impact factors and other citation statistics, says Rosin.

None of these initiatives has yet attempted the multidisciplinary coverage offered by the ISI's database. And given the ISI's dominant market position, it remains unclear whether anyone will. "Even with automation there are high costs of obtaining and organizing a massive amount of journal literature in usable form, and doing quality control," Rosin says. **Declan Butler**

CrossRef

► www.crossref.org

ResearchIndex

► citeseer.nj.nec.com/cs

Open Citation Project

► opcit.eprints.org

Astrophysics Data System

► adswww.harvard.edu

Parity Computing

► www.cparity.com

Declan Butler is *Nature's* European correspondent, and a member of a CrossRef working group exploring search technologies.

papers authored by consortia, rather than a conventional list of individuals. Citations to such papers appear to have been undercounted by the ISI. This came to public attention in January^{3,4}, after *Nature* investigated the suspiciously low citation count for last year's landmark paper on the draft human genome sequence⁵, from the International Human Genome Sequencing Consortium. As it turned out, the ISI was only considering citations to the full list of authors, led by Eric Lander of the Whitehead Institute for Biomedical Research in Cambridge, Massachusetts.

Warped records

Nature is not the first to point out irregularities in the ISI's data^{6,7} (see Correspondence, pages 731–732). Individual researchers have also found errors after probing records for their own publications held by the ISI. Perhaps the largest source of error is the tendency of scientists to make mistakes when citing one another's work. Rather like the attribution of the famous 'lies' quote used to open this article — often associated with Mark Twain but in fact borrowed by the writer from the Victorian-era British Prime Minister Benjamin Disraeli — errors in references mean that citation statistics can misplace credit. Indeed, errors can creep in to every data field recorded by the ISI's database. And as the citation counts creep up, so do the errors as scientists copy them from paper to paper.

Indeed, experts in bibliometric analysis say that, for particularly highly cited papers, it is not unusual to find that variants with, say, an error in the journal volume number have themselves garnered enough citations to beat the vast majority of papers in the ISI's database. Variations in addresses also cause havoc, especially with national, regional and institutional comparisons.

Given these difficulties, and the huge volume of data processed by the ISI — covering some 5,700 science journals — bibliometric experts concede that the company faces an extremely difficult task. The ISI has a quality-control department to clean up the data, and says that the algorithms that match articles have been refined to cope with common typing errors, including misspelt names. But the output quality remains at the mercy of variations in the input. "Nobody knows how accurate the raw data are, but they're certainly not clean, accurate data in the way that some people think," says Ben Martin, a bibliometrics expert at SPRU, the unit for science and technology policy research at the University of Sussex in Brighton, UK.

When specialists such as Martin analyse the ISI's data, they go to great lengths to remove erroneous citations. But anyone who purchases the Essential Science Indicators software can now probe the information in

the company's databases to evaluate institutions, departments and individuals.

Although products such as Essential Science Indicators will allow non-specialists to avoid the mistake of using journal impact factors as a proxy for direct citation counts, many experts are concerned that they will enhance the weight given to citation analysis by evaluating committees. And given that non-experts are generally not aware of the problems caused by errors in the data, this could have serious consequences.

"The data these products produce at the level of the institution or individual can be very 'noisy' and require a considerable amount of cleaning before they can legitimately be used by policy-makers and analysts," says Linda Butler, a bibliometrics researcher at the Australian National University in Canberra. She cites an extreme example: the Walter and Eliza Hall Institute of Medical Research (WEHI), a leading medical facility located in the grounds of the Royal Melbourne Hospital. The hospital appears in the first line of the WEHI address and so collects up to 70% of the citations intended for the independent institute in the subject listings in Essential Science Indicators.



Linda Butler: data can need a lot of cleaning.

The ISI makes such problems clear in the product notes that accompany the software. "But I know from considerable experience that researchers, analysts and bureaucrats will head straight to the data and start playing with them without any real understanding of their quirks or deficiencies," Butler says.

At the same time as launching 'do-it-yourself' citation-analysis software, the ISI has taken a tougher stance on independent research groups that want to use its data. One bibliometrics researcher, who did not wish to be identified, claims that the cost of buying certain ISI data has risen almost fourfold since 1995.

Tibor Braun, a chemist at Loránd Eötvös University in Budapest, Hungary, and editor of the journal *Scientometrics*, calls the era before Thomson bought the company the ISI's "romantic period". Garfield's personal interest in bibliometric research, observes Braun, meant that specialists in the field were given freedom to play with the data. "He pursued many things that were perhaps not essentially business concerns or cost-effective because he owned the company and he was interested in the results," agrees David Pendlebury, an ISI analyst.

The ISI's senior management declined to comment on the company's current business

Nobody knows how accurate the raw data are, but they're certainly not clean, accurate data in the way that some people think.

strategy. But few would dispute that the ISI is acting within its rights in exerting tighter control over its database, which is a valuable asset. For the time being, the ISI retains an effective monopoly in the field of multi-disciplinary citation data — although advances in information technology may yield alternative databases (see 'Pretenders to the throne', opposite).

Mining the business

Braun says that some researchers exploited the freedom they enjoyed in the Garfield era, developing lucrative businesses on the back of the ISI's data. "They saw a lot of people using their database, making products from the information and selling them," says Braun. The ISI's new management may have replaced romanticism with hard-nosed capitalism, says Braun, "but we should not condemn the capitalists".

Garfield confirms that piracy of the ISI's data was always a problem for the company. "From the outset, there were people who felt they could use the data without limit and without cost," he says. "Not the least of these were government bureaucrats who thought nothing of awarding contracts to companies other than ISI to perform studies using ISI data."

Although the ISI's business strategy may be perfectly legitimate, that leaves the bigger question of whether it, and citation analysis more generally, is good for science. Supporters of citation analysis argue that it injects objectivity into decision-making that can otherwise be rife with cronyism. Detractors counter that the practice is so riddled with errors and biases that it can be worse than useless.

As long as citation analysis continues to be used for scientific evaluation, this debate seems sure to continue — and you can cite us on that. ■

David Adam is a news and features writer for *Nature*.

1. Seglen, P. O. *Br. Med. J.* **314**, 498–502 (1997).
2. *Nature Neurosci.* **1**, 641–642 (1998).
3. Cherfas, J. *Science Watch* **13**(1), 8 (2002).
4. *Nature* **415**, 101 (2002).
5. International Human Genome Sequencing Consortium *Nature* **409**, 860–921 (2001).
6. Moed, H. F. & van Leeuwen, Th. N. *J. Am. Soc. Inf. Sci.* **46**, 461–467 (1995).
7. Reedijk, J. *New J. Chem.* **22**, 767–770 (1998).

♦ www.isinet.com

Tightening the purse strings

Science budgets are being trimmed at the European Space Agency. Sally Goodman talks to David Southwood, the man charged with deciding how the money should be saved.

"Nobody does space science who's not an optimist — it's a requirement of the job," observes David Southwood with a wry smile. It is just as well that Southwood, director of science at the European Space Agency (ESA), feels this way. As a space physicist at Imperial College London, he worked on the Cassini mission to Saturn. Now he is responsible for implementing cuts to the very programme that helped to fund ESA's contribution to that project.

Southwood took over as science director last May when astrophysicist Roger Bonnet stepped down after 18 years in the job. He knows ESA well, having chaired the agency's space-science advisory committee and coordinated its Earth observation strategy.

But just months into his new role, Southwood's political paymasters have handed him a budget of 1.869 billion euros (US\$1.62 billion) — 76 million euros less than he had hoped for (see *Nature* **414**, 383; 2001). Science budgets have fallen 15% in real terms over the past six years, he says, and this seriously threatens the breadth of ESA's research.

Recycling project

Southwood now faces the thankless task of deciding how to cut ESA's science programme, for which he has imposed a deadline of June. Money will be saved by reusing hardware developed for other missions. Further collaborations with other space agencies are also on the cards. And a single big mission could still be sacrificed.

But whatever path he chooses, Southwood knows that his decisions will affect hundreds of researchers — many of whom might find it difficult to accept that missions that reuse hardware may win out over others with stronger scientific backing. He says that he is keeping in with his old career by doing a little research when he has the time, but readily admits that the difference now is that "the buck stops here".

Reusing equipment is a tried and tested method of reducing mission costs, but it can complicate schedules. Solar Orbiter, an ESA spacecraft that would observe the Sun from as close as 0.2 astronomical units (1 AU is the distance between the Earth and the Sun), is a case in point. To reduce costs, Solar Orbiter

will use technology being developed to help another ESA spacecraft — BepiColombo — to cope with the high temperatures it will experience on its journey to Mercury.

But BepiColombo will not fly until 2009. So Solar Orbiter's launch must wait until then at the earliest, and could be delayed beyond this date. The industrial teams working on BepiColombo will be disbanded soon after the earlier mission's launch, further complicating matters. Such uncertainty affects the careers of researchers working on the mission.

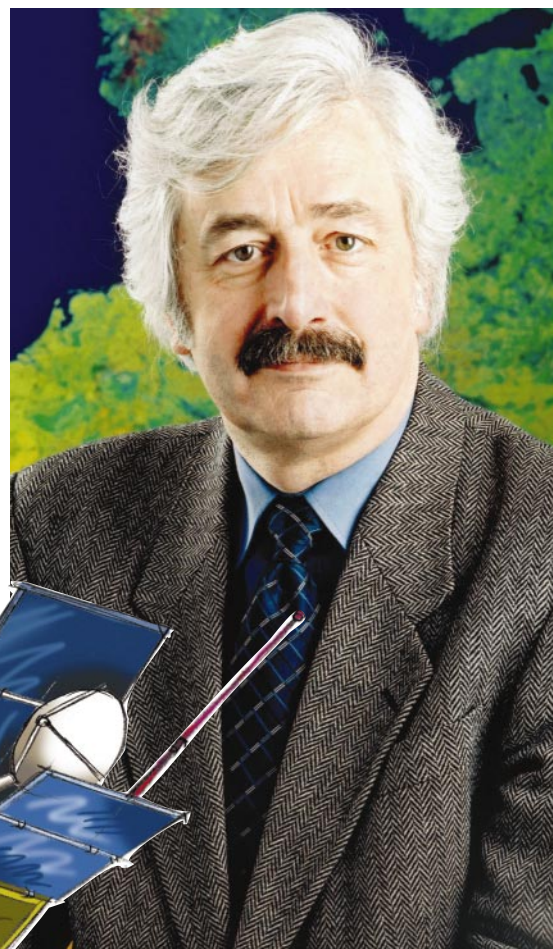
"I have students working on prototypes of instruments for Solar Orbiter," says Ester Antonucci of the Astronomical Observatory of Turin. "But I can't keep them by offering prototypes. To make a career they need to build instruments."

Southwood counters that any decisions on launch dates must incorporate cost considerations. "Scheduling isn't simply by scientific priority," he says. Other missions will have to reuse equipment, he insists. He cites the example of Venus Express, ESA's proposed 2005 mission to Venus, which is designed specifically to use the same type of bus — the aluminium shell that holds the instruments — as the agency's 2003 Mars Express mission.

Mix and match missions

Another solution will be to fill gaps between ESA missions by targeting opportunities in other agencies' projects. European 'space weather' researchers are already working on Stereo, a NASA mission slated for a 2004 launch that will study matter ejected from the Sun. Solar physicists are involved in Japan's 2005 Solar-B probe, which will take high-resolution images of the Sun from an Earth orbit. "The solar physics community can't be dependent on just one agency," says Richard Harrison of the Rutherford Appleton Laboratory near Oxford.

Despite such steps, whole missions may still have to go. After the budget announcement, Southwood publicly stated that GAIA, an ambitious mission to map around 1% of



Outlook sunny? David Southwood's plans for implementing funding cuts could delay the launch of the planned Solar Orbiter (inset).

our Galaxy's stars, was in jeopardy. "I decided to make it clear to ministers, to choose an example, rather than talk in generalities," he explains.

But such talk could turn researchers off the project and become a self-fulfilling prophecy. Gerry Gilmore, a member of GAIA's science advisory group based at the University of Cambridge, UK, already says he would deter students from working on GAIA if the schedule slips, for fear of ruining their careers.

Southwood seems painfully aware of the impacts that his statements can have. "My biggest challenge is keeping the community feeling they've got the best programme for the money," he says. "I don't want disaffected members feeling they've been short-changed."

But cutting costs will inevitably leave some researchers feeling disaffected — a problem only a budget increase is likely to end. In the long term, Southwood hopes that ESA member states will reconsider when they see what cuts mean in practice. "A space agency without a flourishing science programme," he says, "is not fulfilling one of the human aspirations people have for it." ■

Sally Goodman is *Nature's* French correspondent.

♦ <http://sci.esa.int>

P. SEBIROT & C. VIOUX/ESA

The impact-factors debate: the ISI's uses and limits

Towards a critical, informative, accurate and policy-relevant bibliometrics.

Sir—Your Opinion article “Errors in Citation Statistics” (*Nature* **415**, 101; 2002) identified how journal impact factors compiled by the Institute for Scientific Information (ISI) are sometimes included as variables in mathematical formulae and directly influence funding decisions by individual research departments. Such use is inappropriate and counterproductive. In addition, the understandably negative reactions of the scientific community towards this type of use mask the great potential of bibliometric methods.

You also uncovered a particular type of inaccuracy in the ISI's citation rate of a paper by a consortium. This is the tip of an iceberg: we have detected errors skewing results among papers, authors, journals and countries, as well as other sources of error, in a large study we have carried out (details available from H. F. M. at moed@cwts.leidenuniv.nl). Scientists subjected to bibliometric assessment and policy officials using it should be aware of these limitations and problems, so that they can properly evaluate and use it (see Box below).

Bibliometric indicators reflect scientific impact, not quality, and provide useful supplementary tools in the evaluation of academic research, provided that they have a sufficiently high level of sophistication; that their pitfalls are taken into account; and that they are combined with more

qualitative types of information. Their application can stimulate useful discussion among scientists and research managers about publication strategies and research directions; help peer-reviewers to make quality judgements; and enable policy officials and science administrators to raise critical questions about many aspects of scientific activity and to provide insight for policy (funding) decisions.

Individual scientists may wish to assess their publication strategies, or examine the impact of their work on related fields. Managers of research departments may wish to compare the performance of their department with those of competitors and assess their collaboration strategies. A review committee may have the difficult task of assessing a country's performance in a particular field compared with that of others. The dean of a medical faculty may wish to examine whether it can qualify as a ‘research’ faculty. A science minister may wish to assess procedures for submitting proposals for research projects to a national research council. Journal publishers or editors may wish to assess the position of their journals, or find suitable reviewers for submitted manuscripts.

For all these needs, context-specific bibliometric indicators were developed as supplementary tools. Yet an indicator that is useful in one context may be inappropriate

in another. For instance, in a field in which international journals are dominant channels of written communication, journal impact factors are useful measures if calculated accurately. But such measures have no value in assessing individual scientists or research departments. There can be no direct relationship between statistics such as journal impact factors and policy decisions. Such statistics provide indications and category classifications rather than precise measurements. They need to be adjusted and fine-tuned in close interaction with users and with the scientists who are being evaluated, which may require a long development process. Other types of information should also be taken into account.

In our institute's huge analysis of more than 20 million cited references matched to 8 million target articles extracted from the *Science Citation Index (SCI)* and related ISI citation indexes, we found that when data are derived from ‘simple’ or ‘standard’ citation-matching procedures, citation statistics at the level of individuals, research groups, journals and countries are strongly affected by sloppy referencing, editorial characteristics of scientific journals, referencing conventions in scholarly subfields, language problems, author-identification problems, unfamiliarity with foreign author names and ISI data-capturing conventions. The overall number of discrepant cited references is about 7% of the number of citations obtained in a simple matching procedure similar to that applied by the ISI in establishing citation links in the Web of Science and calculating statistics for its newsletter *Science Watch*. Typical examples of strongly affected entities are ‘consortium’ papers; journals with dual volume-numbering systems or combined volumes; journals published in different versions applying different article-numbering systems; and authors from non-English-speaking countries.

This 7% of lost citations skews the distribution of discrepant citations, making some statistics highly inaccurate. For instance, a group of scientists collaborating in a consortium may lose all their joint impact; authors from China or Spain may lose 13% and 8% of their overall citations, respectively; journals publishing combined volumes, such as *Applied Surface Science* and *Physica B*, lose 15–20%. When Spanish or Chinese authors publish the main part of their output in these two journals, the percentage of ‘lost’ citations can easily rise to 25–30%.

Although the ISI is a monopoly supplier

Crucial questions for producers and users of bibliometric statistics

1. Which version of the SCI was used?

The various versions such as the CD-ROM and the Web of Science have different journal coverage, and statistics may differ between versions.

2. How were publication data collected?

Did the scientists subjected to the analysis verify the data? Which variations of author or departmental names were taken into account? Omission of even one highly cited article can substantially distort the results.

3. What percentage of the total publication output is covered by the SCI and included in the dataset analysed?

As a rule of thumb, if this is below 60%, the picture provided by SCI-based statistics may be incomplete.

4. How were cited references matched to target articles?

Did it take into account variations in author names, or discrepancies due to editorial characteristics of articles or

journals? Simple matchkeys may yield highly inaccurate statistics.

5. Do the indicators take into account differences in citation and publication characteristics among scientific fields?

ISI journal impact factors do not take these differences into account and therefore have a limited value.

6. What is the policy question to be answered or problem to be solved?

Indicators are context-dependent and need fine-tuning. Those that are useful in one context may be inappropriate in another.

7. What quality factors are used by evaluators and what is their relative weight?

Evaluators must make this explicit, and not hide behind bibliometric indicators.

8. Do the procedures allow for comments by the scientists subjected to the analysis?

Knowing both sides is indispensable for a proper interpretation of bibliometric statistics.

of the *SCI* and Web of Science, it is not a monopoly supplier of bibliometric statistics derived from these bibliographic information products.

Bibliographic and bibliometric use are two distinct types of use of scientific information, each with its own set of operational and quality criteria. The ISI's information products are primarily developed for bibliographic use. When conducted properly, bibliometrics can unravel relationships that were previously unknown, and put new issues on the political agenda. It can be informative in providing condensed overviews of publication and citation frequencies, and

accurate if proper data-collection procedures are applied.

Anyone confronted with bibliometric statistics derived from the *SCI*, intended to be applied at the level of individuals or research groups or departments, should know the answers to the questions summarised in the Box.

These minimum criteria are crucial for assessment of the accuracy, validity and usefulness of bibliometric statistics.

Henk F. Moed

Centre for Science and Technology Studies (CWTS), Leiden University, PO Box 9555, 2300 RB Leiden, The Netherlands.

Statistics hide impact of non-English journals

Sir — Your opinion article (*Nature* **415**, 101; 2002) provides timely cautions about the errors of *Science Citation Index (SCI)* data compiled by the Institute for Scientific Information (ISI). As Chinese scientific publishers, editors and researchers, we wish to point out that these errors are far more serious for journals published in non-English-speaking countries. The ISI's coverage of scientific journals from these countries is far too limited, and the livelihood of many decent journals has been an unintended casualty.

The ISI misses the fact that non-English-language journals often have alternative names. The *Chinese Journal of Geophysics* — *Chinese Edition* is also called *Diqiu Wuli Xuebao*, *Acta Geophysica Sinica* and *Chinese J. Geophys.* for historical reasons. In 2000, this journal was cited more than 260 times, yet the ISI's *Journal Citation Reports 2001* gives it a total citation score of 13.

The ISI fails to note that many English-language editions of journals published in China have Chinese-edition counterparts, with different contents and domestic and ISSN registration numbers, and should be considered as different journals. Among China's 12 English-language journals indexed by *SCI*, 8 also have Chinese editions. In 1998, for example, 91 of the 147 cited items for papers published during 1996 in the *Chinese Science Bulletin*, a prominent science journal, were incorrect: 52 of the 91 errors were attributions to the bulletin, although these citations were actually to the Chinese edition.

In addition to your reminder to people to use citation statistics prudently, we suggest the ISI should pay more attention to journals in non-English-speaking countries. Even though the ISI should not be held responsible for problems in doing science in developing countries, it can certainly be more accurate in its analysis of

scientific achievements in places such as China, and hence help to promote international scientific communication.

Shengli Ren*, **Guang'an Zu***,
Hong-fei Wang†

**Department of Publication, National Natural Science Foundation of China, 83 Shuangqing Road, Haidian District, Beijing 100085, China*

†Institute of Chemistry, Chinese Academy of Sciences, No 2, 1st North Street, Zhong Guang Cun, Haidian District, Beijing 100080, China

Strange results mean it's worth checking ISI data

Sir — Like the editors of *Nature* (see *Nature* **415**, 101; 2002), we too realize that it is in our best interests to carry out editorial checks of data from the Institute of Scientific Information (ISI). In 1998, *The Lancet's* impact factor, as calculated by ISI, dropped to 11.79 from a previous stable value of about 17 for the previous four years. In 1997, *The Lancet* had decided to divide letters into Correspondence (not counted in ISI's denominator) and Research Letters (peer-reviewed and containing original data, hence coded by the ISI as citable for the denominator). This division increased our number of citable items and thus we attributed most of the drop to this change, an assumption confirmed by the ISI.

Recently, we noted that the number of citable items listed for 2000 was higher (821) than informal calculation would suggest. After hand-coding each issue, we found that 684 items should form the denominator.

Meanwhile, out of editorial interest, we looked at citation data for 1999 in a file purchased from the ISI (based on papers published in 1997 and 1998). As *Nature* did, we also found examples of unexpectedly low numbers of citations for large trials that had only a group name in the author byline compared with those that had at least one named author. The International Stroke Trial (*Lancet* **349**,

1569–1581; 1997), for example, was listed as having one citation in 1999. The ISI confirmed to us that group names are a potential landmine for citation accuracy and said it was in the process of developing a new program to address this issue.

Kathleen D. Hopkins, Laragh Gologly, Sarah Ogden, Richard Horton

The Lancet, 84 Theobald's Road, London WC1X 8RR, UK

See also the News Feature "The counting house" on pages 726–729 of this issue.

Habilitation not just alive in France, but growing

Sir — In your News Feature (*Nature* **415**, 257; 2002), you suggest that the *Habilitation* postdoctoral thesis is "unique to German-speaking countries". Sadly, not so. *L'habilitation à diriger des recherches* is alive and kicking in France, and is an obligation for anyone who supervises a PhD student. Although not usually as large as its German equivalent (mine was only 15,000 words long), the rules governing its size are effectively determined by each faculty. There is a tendency for it to creep up to the size of the old *thèse d'état*, usually more than 100,000 words long, which it was designed to replace in 1988.

Matthew Cobb

Laboratoire d'Ecologie, Université Paris-6, 75005 Paris, France

Getting space camera back on track soon

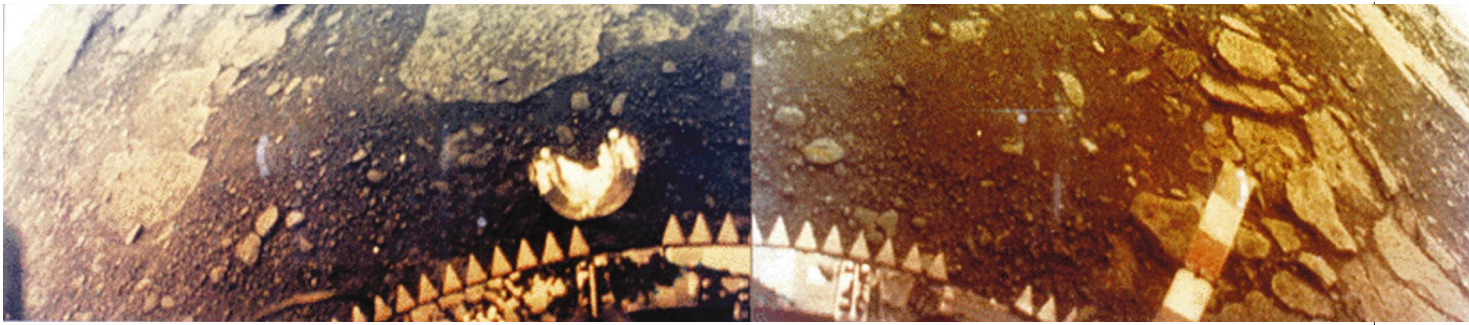
Sir — The headline on your News in Brief "Equipment failure derails space projects" (*Nature* **414**, 835; 2001) does a disservice to at least one of the projects discussed, the NASA–European Space Agency Cassini mission to Saturn. Your article accurately states that Cassini engineers are making progress in fixing haze on images that is thought to result from contamination on the camera's optics or detector. It is, however, incorrect to say that Cassini has been derailed by this problem.

Franklin O'Donnell

Jet Propulsion Laboratory, 4800 Oak Grove Drive, Pasadena, California 91109, USA

Errata In the Correspondence "How scientists can take the initiative in schools" by Mo Afzal (*Nature* **415**, 364; 2002), the citation *Nature* **414**, 1; 2001 should have read *Nature* **414**, 673; 2001.

In the Correspondence "False samples are not the same as blind controls" by L. S. Mills (*Nature* **415**, 471, 2002), reference 5 was incorrect. The correct reference in full is: M. K. Schwartz *et al.* *Nature* **415**, 520–522 (2002).



Narrow horizons in astrobiology

Mission planners must realize that astrobiology is more than a search for life.

Michael J. Drake and
Bruce M. Jakosky

Astrobiology burst upon the world in the mid-1990s, as the excitement about possible life on Mars recorded in the meteorite ALH84001 (ref. 1) and the discovery of planetary systems around other stars² received intense international attention. Was the answer to the question that has been asked ever since humans first became sentient — “Are we alone in the Universe?” — finally within our grasp?

Astrobiology provided, for the first time since the end of the US/Soviet cold war race to the Moon, a deep rationale for the US civilian space programme. What better justification for spending more than \$2 billion a year on space science than to seek the answer to an ancient and profound question? But that goal is not yet being realized. Along the way, the broad and deep concept of astrobiology has, in practice, been reduced to a very narrow interpretation.

Initial promise

Astrobiology, as originally conceived, addressed far more than just the search for life in our Solar System. It involved understanding the current states of planets; how they formed and the nature of their initial states; the evolutionary processes that, over 4.5 billion years, led to their current states; and how those same processes might have operated in other planetary systems. It was about understanding planetary habitability and planetary non-habitability, as well as the actual distribution of life in our Solar System and elsewhere in our Galaxy. (See discussions in ref. 3.)

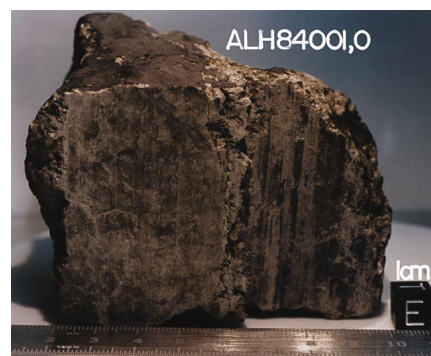
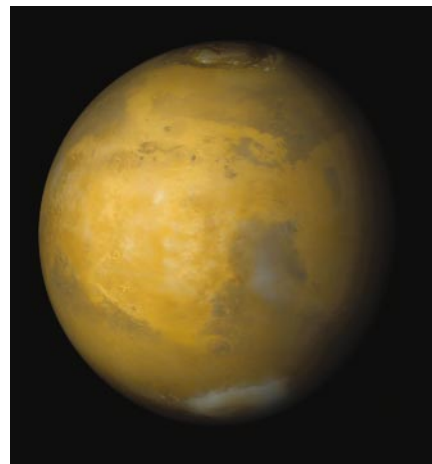
To achieve these goals, we need to know what the ‘building blocks’ were at time zero, before recognizable planets emerged from the disk of gas and dust surrounding the young Sun. The young Sun itself is the product of the collapse of an irregular, cold, slowly rotating cloud of gas and dust that formed from the debris from earlier genera-

tions of stars that lived during the first 10 billion years of the Universe.

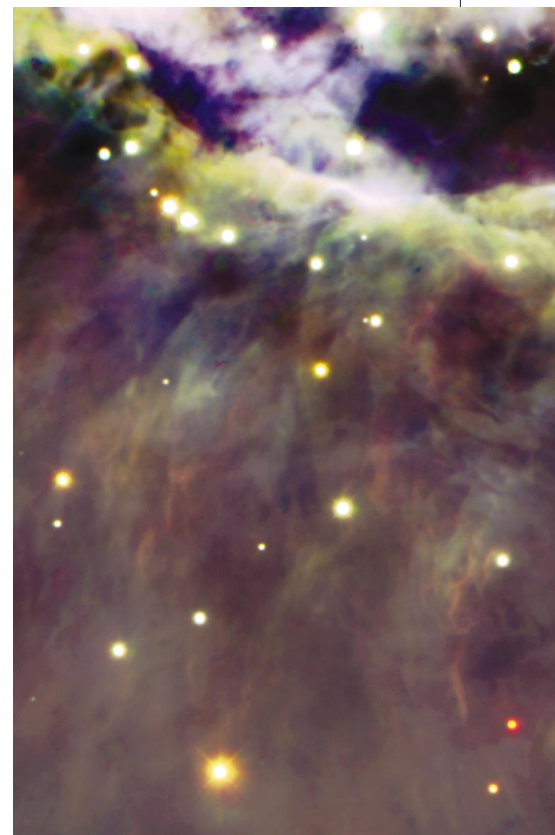
What were the diverse processes of planetary evolution? Why, for example, did Earth, Venus and Mars end up so different from each other? The study of planets and satellites that are hostile to life — as well as investigations of those that might support it — provides insight about the starting conditions and evolutionary processes that cause some of them to be hospitable to life and others sterile.

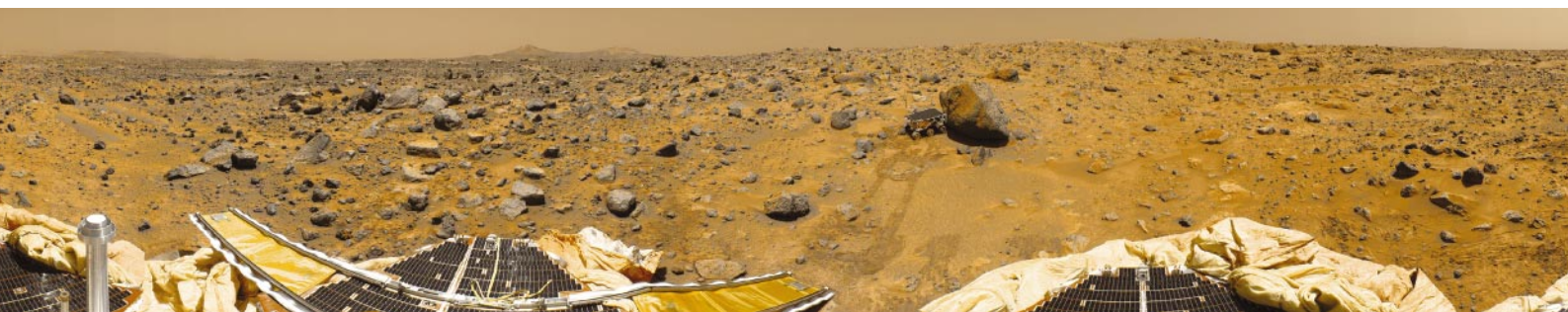
We need to know how oceans and atmospheres form and operate on some planets,

but also why they do not exist on others. We need to find out why some planets have magnetic fields and some do not, the longevity and stability of these fields, and their role in the evolution of atmospheres or in protecting life from hazardous cosmic rays. We need to know the bombardment history of planets to understand whether massive impact events snuffed out early attempts by living things to flourish. We need to understand how planetary systems form — for example, is a giant planet, such as Jupiter, essential for complex life to evolve because it is a ‘cosmic vacuum cleaner’, sweeping up comets that might have



Astrobiology sparked the public imagination for exploring parts of our Universe (right) after hints of life were seen on a meteorite (above left) from Mars (top left). But inhospitable planets such as Venus (top of the page) have been passed over since the Magellan mission was approved in the early 1980s.





▶ had severe impact effects on smaller planets' environments?

Astrobiology requires a rich backdrop of diverse missions, observations, laboratory experiments and theoretical calculations to achieve its promise of understanding the connection between life and planets. This broad picture is needed to understand how these same processes might have occurred in other solar systems, and whether there are likely to be habitable planets elsewhere. In addition to missions to the planets, satellites and smaller objects in our own Solar System, it requires detailed observations of other solar systems and of the characteristics of the planets that we are finding there.

In this context, astrobiology is about much more than just the search for life on Mars or Europa (one of Jupiter's moons). Although some people might expect to find evidence for current or past life on these two planetary bodies, failure to find evidence of life there does not mean that the US astrobiology programme has failed. On the contrary, it is an important scientific result that tells us much about the planetary conditions that allow life to originate or exist and those that are detrimental to life.

Evidence of the absence of life is as important in understanding where and why life exists as finding evidence for life — although, admittedly, the headlines would be less garish.

But the reality is that NASA is implement-

ing an astrobiology programme in an emasculated form. On the one hand, the programme to understand extrasolar planets is developing quite well, with plans for Earth-orbiting telescopes to look for planets around other stars and then to determine their characteristics. But the current emphasis of Solar System spacecraft missions on the search for life, although an important intellectual element of exploration, has led to the effective exclusion of any programmes that do not have the direct and immediate goal of addressing whether life can or does exist elsewhere. The exception is the community-driven Discovery programme of small, low-cost missions, including missions to Mercury (Messenger) and to comets (Contour, Stardust and Deep Impact). The Cassini mission to Saturn, the last of the 'big budget' missions, was put in place before NASA's current mission philosophy took hold.

Restricted view

The executive branch of the US government has restricted new planetary space missions to Mars and Europa, the two most promising places after Earth where life might have emerged or persisted. These places are important targets for exploration, but they certainly do not alone constitute a programme to address the broad goals of astrobiology. Missions to Venus, Earth's twin, appear to have no place in this narrow construction of the field. What can be more important than to know why Venus, so like Earth, is so inhospitable? The trite answer, that Venus is closer to the Sun, is a cartoon approach to the problem. The current approach also does not allow room for missions that have very high priority within planetary science yet have lower priority for astrobiology, such as the Pluto flyby mission, which was funded by US Congressional action after it was left out of the president's budget for the 2002 fiscal year.

As this Commentary went to press, NASA's proposed budget for the fiscal year 2003 cancelled the New Horizons Pluto mission and Europa orbiter, replacing both with the more general New Frontiers programme.

The scientific community has to take its share of the blame for the current narrow emphasis in the interpretation of astrobiology. Few planetary scientists understand the diverse and complex relationships between different components of the system as described above. It has been easier for them

In vogue: Mars has proved to be the main focus of NASA's current efforts in astrobiology.

to take 'astrobiology' as the narrow search for life on Mars and Europa, treating it as separate from the 'true' discipline of planetary science, and then lambast the astrobiology programme for its narrowness. Yet the same community has simultaneously allowed the Solar System exploration programme to be carried along by the public's excitement about the potential for finding life.

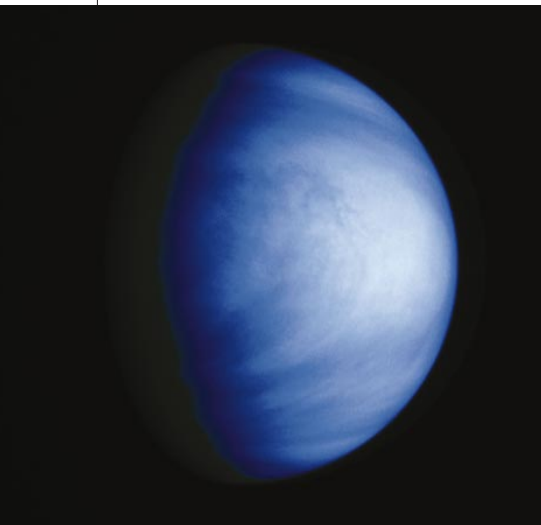
Returning to its roots

Astrobiology is a good idea. When taken in its broader form, it is a unifying theme that resonates with people of all ages, the scientific community and the US government, and its appeal is international. We must break free of the narrow constructionism of astrobiology currently in vogue in the United States. Yes, we would like to know if we are alone in the Universe. But to do that we need to know how solar systems form, how planets evolve, what their interiors are like, how and when oceans and atmospheres form, and how and why environments that are conducive to life emerge. We need to understand what processes resulted in the architecture we find today in our own Solar System, and then how these same processes played out to such different ends in the planetary systems we are finding elsewhere.

To accomplish this, solar-system exploration needs to be balanced and inclusive. A programme of missions, observations, calculations and experiments to address the entire breadth of goals in astrobiology is required. The current Solar System exploration programme is too narrow. Mars and Europa alone do not constitute balanced solar-system exploration, even when the Discovery programme is taken into account. Although our current efforts might, if we are lucky, allow us to answer the question, "Are we alone in the Universe?", they will not address the more important question of "Why?" ■

Michael J. Drake is at the Lunar and Planetary Laboratory, University of Arizona, Tucson, Arizona 85721-0092, USA; Bruce M. Jakosky is at the Laboratory for Atmospheric and Space Physics and Department of Geological Sciences, University of Colorado, Boulder, Colorado 80309-0392, USA.

1. McKay, D. S. *et al. Science* **273**, 924–930 (1996).
2. Mayor, M. & Queloz, D. *Nature* **378**, 355–359 (1995).
3. Astrobiology Insight. *Nature* **409**, 1079–1122 (2001).



Missed out: a trip to Venus could offer clues to the processes that make planets hospitable.

Carry a big stick

A unique figure of the twentieth century gives his version of the facts.

Memoirs: A Twentieth-Century Journey in Science and Politics

by Edward Teller, with Judith Shoolery

Perseus: 2001. 640 pp. \$35, £24.99

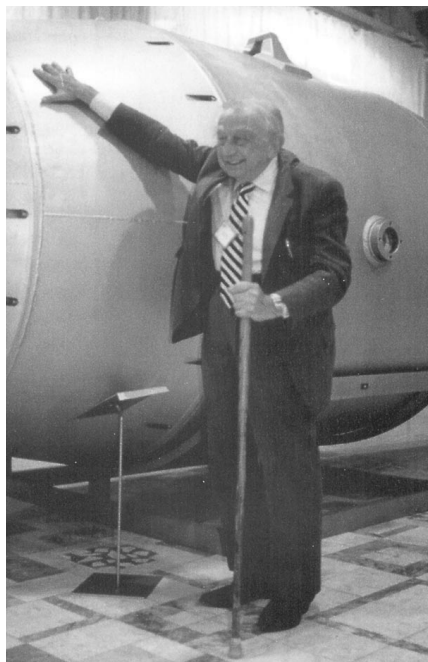
John T. Finn

Edward Teller's hawkish politics and role in the development of the hydrogen bomb leave almost no one ambivalent. As a result, he elicits emotional, polarized reactions like few other scientists. At the White House in 1987, for example, Ronald Reagan introduced him to Mikhail Gorbachev as "the famous Dr Teller". The man whose reforms would later end the cold war responded, "There are many Tellers", and refused the ageing weaponeer's outstretched hand. But Gorbachev was mistaken. Love him or hate him, there is only one Teller, and now — in his tenth decade — he reflects candidly on a remarkable life in science and politics.

Steadfast in his beliefs (and certainly not shy about expressing them), Teller admits to mistakes and possessing his share of that "general human property, stupidity". He also reveals a warmer side to his Dr Strangelove mantle. Part apologia, part anecdotal history of the nuclear age and primer on modern physics, *Memoirs* is a highly readable account of one man's participation in some of the most controversial events of the twentieth century.

Born in Hungary in 1908 and raised in a well-off Jewish family, Teller was a melancholy youth who, by adolescence, had experienced "war, patriotism, communism, revolution, anti-Semitism, fascism, and peace". At eighteen he was off to Germany, where, after a fling with chemistry, he settled as a PhD student in Leipzig with physicist Werner Heisenberg. He became a scientific journeyman upon receiving his degree, working with Niels Bohr and Enrico Fermi and taking a junior faculty position at the University of Göttingen. Despite the ominous rise of the Nazi party, Teller remained strikingly apolitical — and understandably content with being young, brilliant, and at the centre of the new quantum physics.

His first significant notice of world affairs seems to have come in 1933, when he had to flee Hitler's Germany. After stays in Copenhagen and London, Teller arrived in the United States in 1935, where he soon adopted the quiet life of a college professor with "no further ambitions and no interest in changes". But a few years later came the discovery of nuclear fission, and with it the potential for a weapon that would change the world. While listening to President Franklin D. Roosevelt speak in 1940, Teller realized that, as one who



Teller, seen here with two 'big sticks', one of them a hydrogen bomb, was motivated in his war work by the fear that Germany would build a bomb first.

had been able to escape from the Nazis, he had an obligation to "protect freedom" and thus decided to join the Manhattan Project to develop an atomic bomb under the leadership of J. Robert Oppenheimer.

Teller never looked back. He worked diligently on the first nuclear reactor, helped to establish the laboratory at Los Alamos in New Mexico, calculated the effects of a fission explosion, and contributed to the design of the atomic bomb, motivated by the fear that the Germans — under Heisenberg's leadership — would build one first. Ever entrepreneurial (and hawkish), Teller spent much of his time exploring the possibility of developing a more powerful weapon based on a thermonuclear reaction. Although disappointed at being passed over in favour of Hans Bethe to head the theoretical division at Los Alamos, the camaraderie and sense of purpose he experienced made this the "most wonderful" time of his life.

The war won, Teller returned to academia. But when the Soviets successfully tested an atomic bomb in 1949, he lobbied intensely for starting work on a thermonuclear (hydrogen) bomb. Upon President Harry Truman's authorization, he helped to guide the bomb's development and campaigned for a second nuclear laboratory (now the Lawrence Livermore National Laboratory in California). His efforts won kudos from conservatives and the military, but he was opposed by many

others, including Oppenheimer, for advocating a weapon so destructive that its use would cause a vast number of civilian casualties and for contributing to an arms race that would dominate geopolitics. Teller remained (and remains) unmoved by these arguments. When the physics behind the new device was validated by obliterating an islet in the Pacific in 1952, an anxious Teller — far away in California — saw the resultant blip in the output of a seismometer and sent a telegram to Los Alamos that famously proclaimed, "It's a boy".

Teller managed to generate even more controversy in 1954 when he testified against Oppenheimer, whose ties to communists had led to a hearing that would revoke his security clearance and effectively end his career as an influential adviser to the government. Teller claims that he had no doubts about his former colleague's loyalty to the United States, despite his resistance to the hydrogen bomb, and was "certain that Oppenheimer would be cleared". Rather, he felt that Oppenheimer was a "man of unusual emotional needs", and he travelled to Washington only with the intention of demonstrating "the degree to which Oppenheimer's advice [on matters of national security] was wrong-headed". But Teller claims that, moments before testifying, he was shown a transcript of Oppenheimer's testimony in which he admitted to lying to an investigator and falsely accusing a close friend of espionage. Consequently, Teller changed his mind about Oppenheimer, and when asked during the hearing whether he believed his associate to be a security risk, he replied that he "would like to see the vital interests of this country in hands which I understand better and therefore trust more".

The fallout from his testimony was as noxious as that from a nuclear explosion. Teller still feels the pain of being ostracized by much of the scientific community, including many friends. But he remains vocal about his beliefs and sanguine that science can deliver a better world; he has tirelessly promoted nuclear energy, science education, and many other causes. In the 1980s, he again found himself in the cross-hairs when he emerged as a prominent proponent of national missile defence, serving as an adviser to Reagan. Teller drew fire for advocating an expensive system that relied on unproven technology (such as an X-ray laser driven by a nuclear explosion) that some thought would escalate the arms race and lead to the deployment of weapons in space.

Missile defence remains controversial to

this day — much like the nonagenarian author. Unwavering, outspoken and somewhat self-promoting, he admittedly uses *Memoirs* to present his “version of the facts”, but he also writes with warmth, humour and optimism. There is indeed only one Teller. ■

John T. Finn is in the Systems Research Department, Sandia National Laboratories, Livermore, California 94551-0969, USA.

A toast to the genome

Life Script: How the Human Genome Discoveries Will Transform Medicine and Enhance Your Health
by Nicholas Wade

Simon & Schuster: 2001. 208 pp. \$24, £18.99
Carina Dennis

Many of the milestones along the epic journey to sequencing the human genome first came to light in the headlines of newspapers and on television. Rivals from different camps would compete for the bigger

media splash, with the elbowing intensifying in the lead up to the publication of the sequence in *Nature* and *Science* early last year. Nicholas Wade is the *New York Times*’ journalist at the genome frontline; he draws on his own extensive reporting of the subject to bring us a book that recounts the events marking the ‘dawn of the genome revolution’ and illustrates how it lays the groundwork for a new era in medicine.

Wade hits his stride early in the book, portraying both the science and the drama behind the ‘race’, between the international consortium of academic researchers and the private venture Celera, to sequence the human genome. The story doesn’t need any added spice, for it is already peppered with turf wars, public and private mêlée, vendetta and vindication. (Surely the mini-series can’t be far behind.) Notably, Wade tempers the theme of the ‘race’, diplomatically proclaiming no clear winner, as it does indeed depend “on the yardstick applied”. Nevertheless, he does not shy away from lending his own interpretation of events and their significance, and he tips plausible contenders for the Nobel prize, in the event that it is awarded for the sequencing of the human genome.

At roughly halfway, the book shifts to visions of how genomics will transform human health. At this point, the fluidity of the book is displaced by a list of the many tangents that post-genomic research will take as it strives to conquer human disease. Certainly, the examples given are timely — we read of the recent discoveries of genetic variations that underpin diabetes and Crohn’s disease, the design of drugs to specifically intercept the genetic perturbations of cancer, and the molecular basis for why asthma patients respond differently to a widely used drug treatment. This chapter is a *tour de force* of futuristic genomic medicine, but the jolting delivery and propensity for using technical concepts that are not fully explained may leave the non-biologist reader floundering.

For those tempted to put the book aside at this point, I urge you to persevere. For Wade soon resumes his lucid and evenly paced stride in chapters on two topical aspects of modern research: regenerative medicine and immortality. The author traces the quest to find the elusive means of treating stem cells so that they will eventually repair diseased tissues and rejuvenate dilapidated organs, in addition to the scientific ingenuity that could one day extend the human lifespan. Here, Wade is both entertainer and educator, sweeping the reader along with lively accounts of the experiments that are edging us closer to an era of renewable and sustainable health.

The genomics company Human Genome Sciences (HGS), and particularly its chief executive William Haseltine, feature

prominently, more so than the multitude of other commercial endeavours intent on mining the genome for its treasures. In the background of the genome sequencing effort, the HGS has been busily capturing and cataloguing the expressed sequences of the genome, with the focus being on the signalling molecules that form the essential communication network between cells, and which are undoubtedly involved in cancer and other diseases. Wade’s sympathetic handling of the company’s efforts hint that he has great faith in their success. Nevertheless, as he notes, without publication of their findings, the jury is still out on the merits of their claims.

For those desiring a quick snapshot of how medicine is being transformed by genetic knowledge, this book is well ahead of the textbooks, but it will inevitably date as the field streaks ahead at its breakneck pace. A cautionary note is warranted regarding the genetic panacea seemingly idealized in the book. Amplifying the belief of James Watson, a pioneer of the genomic revolution, that the “biggest ethical problem we have is not using our knowledge”, Wade presents a persuasive argument that the fruits of pharmacogenomics will cure all our medical malaises. This is reminiscent of arguments by its advocates that nuclear power will solve our energy crisis. I dare say that it may make our lives better, provided that we can use it without hurting ourselves in the process.

Caution aside, we have a long road yet to travel, but this book is a toast to your genome and to your good health. ■

Carina Dennis is *Nature*’s Australasian correspondent.

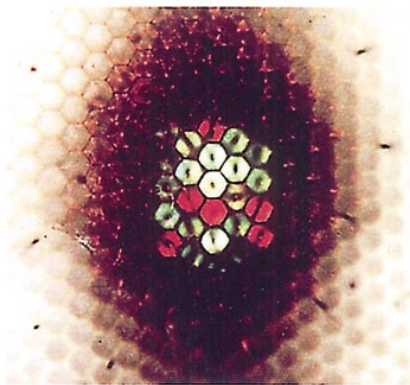
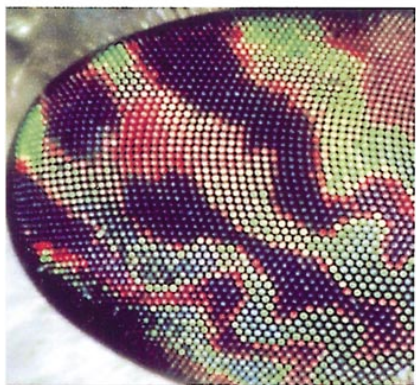
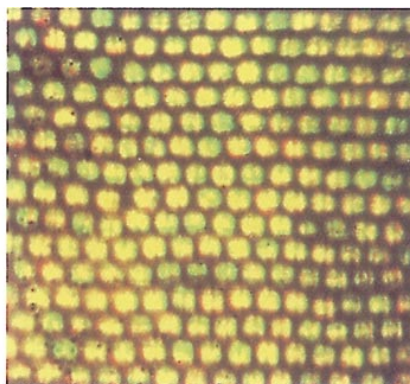
Keeping in sync

Synchronization: A Universal Concept in Nonlinear Sciences
by Arkady Pikovsky, Michael Rosenblum & Jürgen Kurths
Cambridge University Press: 2001. 432 pp. £70, \$100

William Ditto

Have you noticed lately that most technical books make you feel as if you have wandered into the middle of a French art film dubbed in Mandarin? You don’t know whether you care about the issues being addressed or whether it is worth your while to persevere. It excites me when talented scientists break free from the normally turgid prose encouraged by technical journals and write a delightful book that illuminates a subject that is so desperately in need of light. Arkady Pikovsky, Michael Rosenblum and Jürgen Kurths succeed at all levels in writing a definitive exposition of the

Admirable sight: an observer studies the human genome sequence.



Visual virtuosity

The nocturnal wolf spider (top, left; its retina, right) has eight eyes, four of which have a mirror, or tapetum, which reflects light back through the retina. *Animal Eyes* by M. F. Land

and D.-E. Nilsson (Oxford University Press, £24.95, \$45; pbk) discusses such adaptations and many more, including the visual equipment of the horsefly (below, left) and butterfly (right).

phenomenon of synchronization.

Synchronization has been on our tongues since the seventeenth century, when Christiaan Huygens first observed the coordinated movement of the pendulums in his clocks. Today, understanding synchronization may provide the jackhammer we need to penetrate the walls encircling some of our most difficult scientific challenges.

The authors begin with an engaging history of synchronization that whets the appetite. Of course, this history is reassuringly accompanied by black-and-white photographs of cranky-looking scientists — *de rigueur* for any scientific history.

Next comes my favourite section of the book, the plainly titled “Synchronization”. This is written in a way that gives the reader a real understanding of the subject. It leads on to my second-favourite section, the more aggressively titled “What is NOT synchronization” — more scientific books should have a section that clearly spells out misconceptions about their subject.

This attractively laid-out book contains numerous fascinating examples of synchronization culled from the physical and biological sciences. Those that pique the interest include electrical circuits, cardiac pace-

makers, menstrual cycles, circadian rhythms and chaotic synchronization. Attractive illustrations tell the story behind each example and demonstrate an important topic to full effect. There is also a section (often neglected by more reality-challenged authors) called “Detecting synchronization in experiments”, which explores the consequences of when theory meets experiment. This section alone is worth the price of the book.

One-third of the way into the book, I encountered mathematical equations and, amazingly, was sufficiently motivated and had learnt enough from the previous chapters to begin to establish a meaningful relationship with the increasing complexity. I was rapidly guided through the essentials of the synchronization of driven systems, mutual synchronization, noisy synchronization, oscillatory media, phase synchronization of chaotic systems and globally coupled oscillators. These topics are liberally sprinkled with eclectic and interesting examples and with sufficient references to enable the reader to delve further into relevant questions of synchronization in a variety of fields.

The book concludes with one of the best expositions on the synchronization of chaotic systems that I have read. The sections on,

for example, phase (and how it is defined) in chaotic systems, the relationships between unstable periodic orbits embedded in chaotic systems and generalizations of synchronization in complex systems (neglected by most standard treatments) are outstanding.

Clearly, the authors of *Synchronization* have pulled off a very difficult trick, that of writing a book that is both a definitive introduction to synchronization for the casual reader and a definitive text for researchers working in a variety of fields. Still, if I might make a suggestion for the second edition of this book — defy tradition and use pictures of smiling scientists!

William Ditto is in the Wallace H. Coulter Department of Biomedical Engineering, Georgia Tech/Emory University, 315 Ferst Drive, IBB 3306, Atlanta, Georgia 30332-0535, USA.

Words of climatic wisdom

The Encyclopaedic Dictionary of Environmental Change

edited by John A. Matthews *et al.*

Arnold: 2001. 701 pp. £125

Heike Langenberg

Those who have attempted interdisciplinary collaboration know that one of the highest hurdles to overcome can be language. For example, by ‘ventilation’ an oceanographer means an influx of water masses rather than anything to do with wind or air. And words such as ‘vorticity’ are just jargon to non-specialists. These obstacles can make reading literature outside one’s own field of research — as well as talking to colleagues from another department — anything from difficult to hazardous.

The problem is well known from the realm of foreign language. With some training in French but no Latin, at the age of about 15 my sister made an educated guess at the translation of Caesar’s title *De Bello Gallico* and came up with “The beautiful Gaul”. To avoid such potentially embarrassing pitfalls, dictionaries have long served the students of foreign languages. *The Encyclopaedic Dictionary of Environmental Change* aims to support students and researchers of the environmental sciences in much the same way.

But, true to its title, the book is more than just a dictionary. Short paragraphs in understandable language briefly explain the basic concept behind most terms; longer pieces of up to a page are devoted to broader fields such as the “geological record of environmental change”. And references, remarkably up to date and citing a good proportion of work from this millennium, help the interested reader to dig deeper than the book can go. Not entirely true to the title, *The*

Encyclopaedic Dictionary is devoted far more to environmental science and climate reconstruction than to environmental change, but a certain concession to the vogue of the day does no harm and might increase sales.

Of course, even an encyclopaedic dictionary cannot replace a textbook. So if you are thinking of working in the environmental sciences following a degree in another field (mathematics, in my case), you will still need to work your way through the more involved explanations provided in specialist books.

For example, the explanation of the “Coriolis force” spans only eight lines, and does not go into how it works, why it exists or when it is important. But then, when you read a paper outside your own field of expertise and the words “Coriolis force” pop up, you might not want to know all of this anyway. *The Encyclopaedic Dictionary* does tell you that the Coriolis force deflects air flow to the right in the Northern Hemisphere and to the left in the Southern Hemisphere, and that its magnitude is proportional to the velocity of the air mass and the sine of the latitude. It omits to mention that the same holds for ocean currents, which might have been useful. Nevertheless, the explanation should help you to carry on reading your paper with a good idea of what is meant.

The dictionary emphasizes techniques and methods and contains invaluable accounts of the many ways of constructing past, present and future climates from terrestrial and marine archives as well as from models. Given that most of our knowledge of past climate is based on the interpretation of a wide range of complementary proxies — records such as tree rings — which give indirect information, this is perhaps the most important use of the book. Modellers can learn what speleothems can tell us, and soil scientists obtain an understanding of how foraminifera can record climate.

The coverage seems to be pretty complete; at points, the editors even go a little overboard. For example, 4.5 pages of entries with the prefix “eco” (including “ecofeminism”, unknown even to the *Encyclopaedia Britannica*) are not strictly necessary, and instead of three terms starting “geoeco” (geocology, geocosphere and geocosystems), a single one would have been enough. On the other hand, three entries for “hotspot” (“in biodiversity”, “in geology” and “in remote sensing”) are utterly justified and very useful.

All in all, *The Encyclopaedic Dictionary of Environmental Change* helps to make sense of the babel that can be so characteristic of the literature and conferences in the environmental sciences and gives excellent support when reading across the disciplines. But beware: if, like me, you enjoy following up cross-references, this book may bind your attention longer than you had intended. ■
Heike Langenberg is a physical sciences editor at Nature.

Science in culture

Godwin's galaxies

The illustrations in Stephen Hawking's *The Universe in a Nutshell*

Martin Kemp

Illustrations accompanying popular science are rapidly becoming a genre of art in their own right. The most valiantly ingenious are those in books by theoretical physicists and mathematicians who take us on dizzying voyages through multi-dimensional space. The classic dilemma of how to represent figures of more than three dimensions on a flat surface remains the greatest technical challenge, allied to the task of visualizing the kinds of space that evolution has ill-equipped us to ‘see’.

The latest contender in the quest to portray the strange world of theoretical physics is the artist, illustrator and author Malcolm Godwin, whose works are included in such prestigious collections as the Museum of Modern Art in New York and the Tate Gallery in London. Until the 1970s he worked as a sculptor, using translucent materials to explore the representation of complex spaces. As an author-illustrator, he has been responsible for books such as *Who are You? 101 Ways of Seeing Yourself* (Carroll & Brown, 2001), *The Lucid Dreamer* (Simon & Schuster, 1994) and *Angels: An Endangered Species* (Simon & Schuster, 1990).

Having collaborated with Stephen Hawking on *The Illustrated Brief History of Time* (Bantam, 1996), he has now provided the extraordinary suites of images for Hawking's *The Universe in a Nutshell* (Bantam, 2001). Although credited in person only at the very end of the ‘Picture Acknowledgements’, his depictions occupy at least as much space as Hawking's text, and bear a very substantial part of its communicative burden.

Hawking began by signalling where he thought illustrations should be provided, initially envisaging many more than could eventually be included. Godwin, for his part, provided twice as many as were eventually used, and each design was either accepted, returned for revision or rejected by the author.

The range of illustrative techniques is remarkable. There are photographs and images of people, literal depictions, mnemonic representations of objects (such as clocks and guns), poetic evocations of concepts, suggestive analogies and metaphors, representations of two- and three-dimensional space, and illusionist tricks to conjure up visions of multi-dimensional space, as well as the diagrams, graphs and formulae expected in a physics book. Conventional figures are hugely outweighed by those that use visual evocation to depict concepts that cannot be represented in a direct manner.

Not the least of the problems is the disjunction between the language of theoretical physics and that of the illustrator. This applies literally, in that a term such as ‘axis’ has a



Brane new worlds: Godwin's depiction of two adjacent brane worlds, one of them our own.

different meaning for an artist than for a theoretical physicist studying black holes. It also applies more broadly to the way that physicists exploit a quasi-diagrammatic and mathematical language that is internal, part of their own conceptual world. Indeed, Hawking does not demand that his mathematical models correspond to physical reality as experienced in any obvious sense. As a self-declared positivist, the existence of extra dimensions has meaning for him only insofar as the mathematical models of n -dimensional space serve as good descriptions of the Universe.

An excellent example of how Godwin has responded to something that cannot be envisaged as straightforwardly ‘real’ is his illustration of the influence of one ‘brane world’ on another. In brane theory, largely the creation of Hawking's colleague Paul Townsend, a brane (as analogous to ‘membrane’) is defined as an object that has a variety of dimensions. Thus, a p -brane possesses length in p dimensions, whereas a 1-brane is a string, a 2-brane is a membrane, and so on.

The illustration evocatively shows two adjacent ‘brane worlds’, our own and another. The nearby brane world is not visible from ours because light is trapped within the dimensions of each brane. However, gravitational forces are envisaged as operating across the gap, and the existence of the mass of the nearby ‘shadow brane’ can only be detected through its gravitational effect on the behaviour of objects in our own Galaxy. The orbital paths of our stars are thus literally overshadowed by sources that are irredeemably dark to us.

The reader, perhaps struggling to understand when told how brane theory is part of M-theory, might turn to the glossary. This duly informs us that M-theory “unites all five string theories, as well as supergravity, within a single theoretical framework, but which is not fully understood”. A tough job indeed for the illustrator. ■
Martin Kemp is in the Department of the History of Art, University of Oxford, Oxford OX1 2BE, UK.

MALCOLM GODWIN

Innateness

Reconciling 'instinct' with biological reality may require a recasting of evolutionary metaphors.

Barbara C. Scholz

In its everyday use, the word 'innate' says something about origins. If we claim some attribute is innate (or inborn, or inherited), we are saying that it originates internally — its cause is located inside the organism. What is not innate originates in the external world. When behaviours or capacities to behave are innate, the word 'instinct' is applied. If a behaviour or capacity is not instinctive, then its origin must be outside the organism. The metaphor has no room for the complexities of ontogenesis. Saying that Raskolnikov had a criminal instinct labels his criminality as having an internal genesis, but does not explain how it developed, because everyone agrees that instincts can be overcome (a left-handed child can be taught to use his or her right hand instead).

The notion of 'innate ideas', which is evident in philosophy from Plato to Leibnitz and beyond, divides ideas into those that spring from within and those that are caused by anything else. The idea thus amounts to little more than an opposition to learning theory — to call an idea or principle 'innate' is to claim that it is unlearned. The folk-psychological metaphor of innateness exhibited in classical philosophy can perhaps be forgiven for its explanatory failure, but science cannot. Do terms such as 'innate' and 'instinct' embody at least the beginnings of a coherent explanatory concept? Or do they occur in contemporary psychological and biological literature as fossilized folk labels with no explanatory function?

We see the concept of instinct used with at least some explanatory force when Darwin wrote, in *The Descent of Man*, that instincts are "gained step by step, through the variability of the mental organs and natural selection, without any conscious intelligence on the part of the animal during each successive generation". He was saying that, as long as there is variation of heritable propensities for behaviour within a population, the distal causes of instincts are environmental. This does more than label the source of instinct. It provides an explanation of how innate behavioural capacities arise: they are the result of selective pressures on a population. The metaphor suggested by the prefix 'in-' of 'inborn', 'instinct' and 'innate' is thus attenuated somewhat.

But there is also a weakening of the folk-psychological idea that what is innate is unlearned; in fact, this idea does not really survive at all in the darwinian view. Darwin

wrote that he was "very far from wishing to deny that instinctive actions may lose their fixed and untaught character, and be replaced by others performed by the aid of free will". He noted in *On the Origin of Species* that 'instinct' is not definable, and does not denote a natural class. Instead, 'instinct' represents a cluster concept — some, but not all, instincts are unlearned.

It is not going too far to say that Darwin explains the origin of instincts at the cost of being unable to characterize what is being explained. Instinctive behaviours no longer clearly originate inside the organism, nor are they uniformly unlearned.

By the late nineteenth century, 'instinct' was being used by psychologists to label a vast collection of behaviours and capacities; one disgruntled critic counted 280 examples. Someone had to try to make a science out of this burgeoning catalogue. That was the task taken on by the anti-behaviourist Konrad Lorenz — disregarding Darwin's warning that instincts are not a natural class. Lorenz's theory of instincts aimed to identify genetically determined, species-specific traits that can be explained by natural selection. As far as he was concerned, the internal origins of instincts are genetic, and this fact alone can explain how individual organisms come to develop stable, developmentally fixed behaviours. They would therefore be manifest in isolation experiments, which were supposed to show what an organism did when deprived of any opportunity to learn a behaviour through interaction with an external environment. With Lorenz, then, the internal/external dichotomy of folk-psychology returns in force. But the underlying assumption that there is a vast difference between instincts and all other behaviours remains unexplored.

Lorenz was widely criticized for his failure to understand the relationship between genes and behaviour, and for simply labelling this relationship "maturation". Early critics such as Daniel Lehrman complained that not only did Lorenz misconceive organism-environment interactions, but more importantly, he used the concept of maturation as a way of ignoring developmental complexity.

It would seem, then, that if the folk-psychological idea of what is innate is to be fitted into a biologically grounded psychology, it will have to be based on a dichotomy between genetic information on the one hand and everything else on the other.

The recent trend known as developmen-



Criminal creation: the misdeeds of *Crime and Punishment's* Raskolnikov (here in a 1935 film) were probably born of both 'instinct' and society.

tal systems theory (DST) rejects such dichotomizing. In fact, it rejects the entire internal/external distinction that is endemic in folk-psychological ideas of innateness, and recommends the abandonment of such unhelpful metaphors. In the DST view, nature/nurture disputes cannot be resolved by research into how much must be attributed to internal factors and how much to external ones. Such investigations merely rehash the dispute. It is not news that organisms select their environments, and that environments are inherited by organisms. Nor should it be news that focusing obsessively on which parts are innate and which are not obscures understanding of developmental processes. What is news is that dichotomous metaphors must be dissolved by recasting our conceptions of heritability and natural selection to include both what is within and what is without.

For the proponents of DST, in other words, Raskolnikov was a criminal neither solely because of a criminal instinct within him, nor solely because the environment of St Petersburg criminalized him. Yet somehow, he developed into a criminal, nonetheless. The complex, web-like processes of that development could be a profitable subject for science as well as novels. ■

Barbara C. Scholz is in the Department of Philosophy, San José State University, San José, California, USA.

FURTHER READING

Oyama, S., Griffiths, P. E. & Gray, R. D. *Cycles of Contingency: Developmental Systems and Evolution* (MIT Press, Cambridge, Massachusetts, 2001).
Oyama, S. *Evolution's Eye: A Systems View of the Biology-Culture Divide* (Duke Univ. Press, Durham, North Carolina, 2000).

Secret life of genes

Günter Theißen

'Homology' is one of the most important terms in biology. Features are homologous if they share a common evolutionary origin — for example, bat wings and bird wings are homologous as tetrapod forelimbs, but they are not homologous as organs of flight. The definition of homology has changed over time and, although still unresolved, functional similarity has never been part of it (bat wings and bird wings certainly have similar functions).

When homology is applied to genes, at least two fundamentally different subclasses must be distinguished: paralogy, the relationship between genes that have originated by gene duplication; and orthology, which refers to genes that originated by speciation. Phylogenetic reconstructions of organisms created using information from the nucleotide sequences of genes require orthologous, rather than paralogous, genes, so the distinction between these two gene classes is important for practical reasons. More fascinating is the observation that orthologues and paralogues usually have

very different evolutionary fates. Orthologues often take over the function of the precursor gene in the species of origin and thus tend to be conserved. In contrast, young paralogues have redundant functions, which is an evolutionarily unstable situation. Thus, in the long run — with a few exceptions — paralogues either diverge functionally, or all but one of the versions is lost.

Although the theoretical definition of orthology seems straightforward, the term is misused or misunderstood in many ways, particularly in developmental biology and genomics. An example is the romantic, but incorrect, idea that every gene in a species' genome has exactly one orthologue in the genome of another species. Genes can be lost during evolution, meaning that a given gene may or may not have surviving orthologues in other species. Moreover, gene duplications can follow a speciation event, generating orthologous 'clades' of paralogues.

Orthology between individual genes does not therefore exist; rather, one-to-many or many-to-many orthologous relationships are formed. The situation is further complicated by the fact that gene duplication, gene loss and speciation can be frequent events in the history of a group of organisms. Thus, complex gene relationships are established which cannot be described in simple terms.

A related error is the naive assumption that a gene with the greatest sequence similarity to one from another genome is 'the' orthologue of that gene. But in view of gene loss and other events, the most similar genes in two species' genomes may be paralogues, or not homologous at all. Thus, even a comparative analysis of entire genome sequences does not guarantee the identification of orthologues.

In my view, the most serious misconception is to confuse orthology with functional similarity, or to use functional similarity as the decisive criterion by which to identify 'the' orthologue from a group of homologous genes. As a subclass of homology, orthology cannot be identified by functional similarity. Functional similarity may reflect orthology, but it could also be the result of convergent evolution. And it is clearly possible for orthologues to diverge functionally. Hence, orthology and functional equivalence must be tested independently, by phylogenetic reconstruction and comparative functional studies, respectively.

It follows that a reconstruction of gene phylogeny, and its superimposition on the

Orthology

The relationship between genes that originated through speciation is one of the most widely misunderstood — and misused — concepts of evolutionary biology.

clarified evolution of the taxa involved, is the only valid way in which true orthologues can be identified. As this is a laborious task, and the available data are often too limited to provide conclusive results, proven (rather than putative) orthologues are as rare in the literature as diamonds in bare rock. On the other hand, experimentally untested claims about orthology seem to be as numerous as grains of sand on a beach.

Of course, the definition of a term may change if this reflects an improved understanding — a good example is the definition of homology, which, unsurprisingly, has taken common ancestry to be the decisive criterion only since Darwin's time. What is bad about using the term 'orthology' to denote the relationship between homologous genes that are functionally equivalent in different species is that this habit undermines a better understanding of some of the most fascinating aspects of nature. Because the complexity and diversity of living beings are largely encoded in their genes, the mechanism and timing of how genes originate and evolve is of the utmost importance.

For a precise description of gene evolution, we need a controlled vocabulary that distinguishes strictly between gene genealogy and gene function. Only then can we properly describe — and eventually understand — the conservation and change of gene function during evolution, and the consequences of these processes for the phenotype. Outside evolutionary biology, 'orthology' is becoming a fashionable, yet fuzzy, buzzword. But terms are tools — care is required to ensure they are not blunted by over-frequent use. ■

Günter Theißen is at the Genetics Lehrstuhl, Friedrich Schiller University, Philosophenweg 12, D-07743 Jena, Germany.

FURTHER READING

Fitch, W. M. *Syst. Zool.* **19**, 99–113 (1970).
Fitch, W. M. *Trends Genet.* **16**, 227–231 (2000).
Graur, D. & Li, W.-H. *Fundamentals of Molecular Evolution* (Sinauer, Sunderland, Massachusetts, 2000).
Tatusov, R. L., Koonin, E. V. & Lipman, D. J. *Science* **278**, 631–637 (1997).

CAGC
ACCA
CAGC
ACC
CA
AC



AGCG
ACCAU
GCG
CAU
GCG
CAU



Beyond the kitchen sink

Roger C. Newman

If you look closely, stainless steel is not actually 'stainless' — it does corrode. Corrosion is known to develop near sulphide impurities in the metal, but not, it seems, in the way we once thought.

Nowadays we hear so much about new or 'advanced' materials that we rarely reflect on the virtues of mature metals technology. A good example is stainless steel — 'stainless' because it is more resistant to rusting than ordinary steel. Most modern stainless steel is a ductile alloy containing iron, chromium and nickel, a technology pioneered early in the twentieth century by the German company Krupp. At the same time, Harry Brearley in Sheffield introduced hard iron–chromium–carbon grades for cutlery. All these materials have very good corrosion resistance. But, on close inspection, stainless steel does suffer corrosion, and it usually begins at sites in the metal where there are sulphide impurities. In this issue (*Nature* **415**, 770–774; 2002), Mary Ryan and colleagues present a new explanation for how the presence of sulphides can lead to corrosion.

Stainless steel resists corrosion because its chromium content (typically 13–25%) is enough to transform the surface oxidation process in wet environments. If we abrade a piece of stainless steel and dip it into water, a protective, chromium-rich oxide film forms over its surface within a fraction of a second. Electrochemical experiments show that, after a few weeks, the corrosion rate of the stainless steel is about 10 nanometres per year — 10,000 times slower than that of ordinary carbon steel. At this rate, a plate of stainless steel 1 cm thick should last a million years (it would actually last much longer than that — research related to nuclear waste containment has indicated that the corrosion rate falls off indefinitely over time, following a $1/t$ law).

There was no stainless steel a million years ago, but the material has been used in wet environments for more than 80 years — and not just for cutlery and kitchen sinks. In 1929, the seven-storey pinnacle of the Chrysler Building in New York was faced with stainless steel (Fig. 1). The steel has been cleaned only two or three times since, and shows few signs of corrosion, despite the urban environment. But if we were to examine the metal surface with a microscope, we would find small pits caused by intense local corrosion. Closer study would reveal residues of sulphur in many of these cavities.

The pitting is the result of the combined action of aqueous chloride ions from the environment, and manganese sulphide par-



Figure 1 Stainless character. The upper floors of New York's Chrysler Building are clad in stainless steel. Despite more than 70 years' exposure to the elements, there is no obvious corrosion. But on page 770 of this issue, Ryan *et al.* show how chromium depletion initiated by sulphide impurities makes even 'stainless' steel susceptible to corrosion.

ticles — known as inclusions — in the steel. Sulphides are an unavoidable contaminant in the steel-making process and sometimes congregate as liquid steel solidifies. They are known to be a major factor in corrosion, not so much for atmospheric exposure but more for immersion in saline environments, where pitting can penetrate several millimetres of steel in a few weeks or months. Alloying elements such as molybdenum can be added to the steel to provide adequate corrosion resistance in saline environments (such as for medical implants in the human body), but these are expensive.

During the process of pit formation, the manganese sulphide dissolves to form corrosive entities such as hydrogen sulphide — at least, that's what I thought until I read the paper by Ryan and colleagues. Their work is an impressive analytical achievement, based on a technique (secondary ion mass spectrometry) that has so far found most of its applications in microelectronics. According to these authors, pits do not initiate because of chemical corrosion caused by the inclusion; rather, they start in a narrow, chromium-depleted region next to the inclusion. Using a grade of steel 'resulphurized' so as to have a suitably high inclusion content, Ryan *et al.* find that the bulk

chromium level of 17–18% is reduced to just 12–14% in a few hundred nanometres around a manganese sulphide inclusion. The inclusion is enriched in chromium by a corresponding amount. The dip in chromium content around the inclusion weakens the alloy's protective oxide layer, making it vulnerable to corrosion.

If the sulphide impurities could be removed from stainless steel, the corrosion problem in saline environments would be alleviated. Over the years, much effort has been devoted to reducing the amount of sulphur in stainless steel. When the Chrysler Building was constructed, stainless steel was made by melting chromium and nickel with traditional high-sulphur pig iron, so the sulphur content of the stainless steel was probably at least 0.05%. But modern technology, which uses oxygen to remove carbon from molten stainless steel, has the benefit of removing most of the sulphur as well. Once most of the carbon has been oxidized, any oxidized chromium needs to be recovered by reduction to the metallic state. This is brought about using a lime-rich slag, containing the reactive elements silicon and aluminium, which also scavenges sulphur from the liquid steel.

Nowadays the task of sulphur removal

RICHARD HAMILTON SMITH/CORBIS

has been made easier by the generally lower sulphur content of carbon-steel scrap that is used to make stainless steel: sulphur levels of 0.001–0.003% are routinely achieved. For premium products such as medical prostheses, further reductions in impurity levels can be achieved using secondary remelting and refining. These use techniques developed for the manufacture of gas-turbine blades, which have to resist creep (gradual elongation under stress).

But very low sulphur content has a surprising disadvantage: the steel is unsuitable for components that require extensive machining. Here the lubricating effect of manganese sulphide is needed, so it is common to resulphurize the steel to a level of 0.3% sulphur. But such materials can have poor corrosion resistance, so a compromise may be preferred with a controlled sulphur level (around 0.02 to 0.03%) that permits reasonable machinability. Welding is also easier with some sulphur content.

One implication of the research by Ryan *et al.* is that we might be able to improve the corrosion resistance of stainless steel by unconventional heat treatments, designed to diffuse chromium back into the depleted zones. Modifications in alloy composition are also feasible, although these might be expensive, or might affect the mechanical properties or weldability. In this respect, the ductile Fe–Cr–Ni grades represent a tougher challenge than the less ductile Fe–Cr grades, to which titanium can be added to ‘tie up’ the sulphur, preventing the appearance of manganese sulphide inclusions.

We should also reassess the technology for treating stainless-steel surfaces. At present, many products are treated with acid and then brushed with hard polymer brushes or mops (many readers will have both the steels and the brushes in their kitchens). Some products are ‘bright annealed’, which involves heating in a mixture of hydrogen and nitrogen. Cold rolling can be used, before or after these treatments, to improve the surface, or to introduce a pattern.

At the microscopic level, little is known about the behaviour of near-surface manganese sulphides during all this processing. But following on from the work of Ryan *et al.*, techniques of high-resolution microscopy and analysis, backed up by new electrochemical approaches to corrosion monitoring, could be used to control corrosion. There is no reason why the most basic grades of stainless steel should not find wider application — helping to conserve scarce metal resources — once the effects of impurities are understood and controlled. ■

Roger C. Newman is at the Corrosion and Protection Centre, University of Manchester Institute of Science and Technology, PO Box 88, Manchester M60 1QD, UK.
e-mail: r.c.newman@umist.ac.uk

Biogeography

A case of dispersing chameleons

Olivier Rieppel

Chameleon species are widely distributed and provide a test for some basic principles of biogeography. The latest analysis supports the idea that they have dispersed from Madagascar on several occasions.



PETER OXFORD/BBC WILD

For most people, chameleons are lizards that live in trees and are known for their ability to change colour. Others know that these lizards have a laterally flattened body and a prehensile tail. But for Raxworthy *et al.* (page 784 of this issue¹), chameleons have provided a test of some basic principles of how related species of animals come to be geographically distributed in the way that they are.

Almost all chameleons live in Madagascar and Africa, but some species are found in southern Europe and the Middle East, and others in India, Sri Lanka and on islands such as the Seychelles and the Comoros. Raxworthy *et al.*¹ have produced the first broad-based analysis of chameleon inter-relationships. Their results suggest that chameleons dispersed several times out of Madagascar and migrated across the ocean to achieve their present distribution. This conclusion is surprising, because it implies that organisms that are highly adapted to life on land have managed to travel across the sea. It also runs counter to a trend in biogeography, in which models that invoke plate-tectonic movement are preferred to ‘dispersalist’ explanations such as this.

An early distinction was drawn between the widespread, larger and predominantly arboreal chameleons, and the geographically restricted, smaller and mainly ground-dwelling dwarf chameleons. The dwarf

Figure 1 Parson's chameleon, one of the species in the analysis by Raxworthy *et al.*¹

species were later assigned to two separate genera, one from mainland Africa and the other from Madagascar. But beyond these basic divisions, the evolutionary history of chameleons has proved notoriously intractable. Debate has centred on two issues. Are the ground-dwelling dwarf chameleons primitive, or are they descended from arboreal ancestors? And did chameleons originate in Madagascar, or on the African mainland? The fossil record of the group is poor, going back only to the Miocene (no more than 26 million years or so)^{2,3}, and offers no clue to its evolutionary history.

During the late 1970s and early 1980s, there was a reformation in systematics that led to the acceptance of cladistic methods in reconstructing evolutionary histories. As a result, the hunt for actual ancestors was replaced by a search for sister-group relationships — that is, for groups that share a hypothetical common ancestry. At the same time, work in biogeography that was preoccupied with the search for centres of origin, and dispersal from those centres, was challenged by ‘vicariance biogeography’, in which the aim is to explain the geographical distribution of organisms in the light of past geomorphological events. Plate tectonics and consequent effects such

as mountain-building are seen as ways in which the population of a species may be split into two or more. If these events create an insurmountable barrier — as an opening ocean presumably would for chameleons — the subpopulations will pursue separate evolutionary trajectories. Eventually they evolve into separate species, subsequently, perhaps, producing separate evolutionary radiations of species in their different areas.

Vicariance biogeography, then, seeks to match sister-group relationships, as revealed by cladistics, with past geomorphological events. With their occurrence on mainland Africa, Madagascar, India and several islands, some of which have a volcanic origin, chameleons attracted the interest of biogeographers because of their suitability as a test case. India and Madagascar started to break away from Africa in the mid-Jurassic period (165 million years ago), before they themselves separated during the Late Cretaceous (88 million years ago) — long before the first appearance of chameleons in the fossil record, and pre-dating the age of the chameleon lineage as determined by the molecular clock¹.

The first cladistic treatment of chameleons presented a phylogeny that perfectly matched a vicariant explanation of their distribution in terms of plate tectonics⁴. A later analysis⁵ found relationships that were only partially reconcilable with a strictly vicariant pattern of chameleon distribution. But the significance of these early studies was limited because not many species were included, and only a few characters were used to analyse their relationships.

Raxworthy *et al.*¹ have now produced the most complete phylogenetic analysis of chameleons so far. It is based on a large number of species (52, one of which is shown in Fig. 1) and characters (644), both anatomical and molecular, and is better resolved than the earlier attempts. Dispersalist explanations in biogeography had fallen from favour, partly because independent geological evidence can be brought to bear on vicariant models. But the analysis of Raxworthy *et al.* is incompatible with a vicariant explanation for chameleon distribution in terms of plate tectonics. It implies instead that there have been several dispersals from Madagascar, and confirms that dispersal is important in determining biogeographical patterns.

Similar explanations, invoking dispersal, have been applied to the distribution of other organisms from Madagascar and elsewhere. One example involves freshwater fishes known as cichlids⁶; another concerns several groups of terrestrial mammals. The occurrence of these groups in Madagascar has been seen as paradoxical because — as for chameleons — their origin postdates the separation of Madagascar from Africa, and because such animals are considered to be poor dispersers

across oceans⁷. This observation generated research suggesting that there was a land bridge across the Mozambique Channel between the mid-Eocene and the Early Miocene (45 million to 26 million years ago)⁷.

How chameleons managed to disperse across the ocean must remain a matter of speculation. There may have been active migration to the African mainland, as the hypothesized land bridge may have overlapped in time with the earliest diversification of chameleons. Dispersal to other areas, particularly to volcanic islands, probably happened passively, by rafting on trees uprooted by storms, or on pieces of land detached by coastal erosion.

Raxworthy and colleagues' hypothesis¹ is of general interest and will provide the

impetus for further research. The challenge is now to investigate other groups of organisms with a geographical distribution similar to that of chameleons to see whether some regularity underlies their phylogenetic and biogeographical patterns.

Olivier Rieppel is in the Department of Geology, The Field Museum, 1400 South Lake Shore Drive, Chicago, Illinois 60605-2496, USA.

e-mail: rieppel@fmnh.org

1. Raxworthy, C. J., Forstner, M. R. J. & Nussbaum, R. A. *Nature* **415**, 784–787 (2002).
2. Estes, R. *Encyclopedia of Paleoheterology*, Vol. 10A (Fischer, Stuttgart, 1983).
3. Rieppel, O. *et al.* *J. Herpetol.* **26**, 77–80 (1992).
4. Klaver, Ch. & Böhme, W. *Bonn. Zool. Monogr.* **22**, 1–64 (1986).
5. Rieppel, O. & Crumly, C. *J. Zool.* **243**, 351–380 (1997).
6. Vences, M., Freyhof, J., Sonneberg, R., Kosuch, J. & Veith, M. *J. Biogeogr.* **28**, 1091–1099 (2001).
7. McCall, R. A. *Proc. R. Soc. Lond. B* **264**, 663–665 (1997).

Earth science

Flow and fabric deep down

Karen M. Fischer

Conditions in Earth's mantle can be inferred from seismic-wave velocities which reveal areas of anisotropy. Evidence emerges of such an area at mid-mantle level, associated with a subducting tectonic plate.

Imaging the Earth's interior is a challenging business. Drilling can only scratch the surface, so much of our information has to be inferred from the behaviour of seismic waves. Convection processes in the mantle — the nearly 2,900 km of material that lies between core and crust — produce deformation and mineral fabrics that have different physical properties in different directions. This elastic anisotropy creates anisotropy in the velocities of seismic waves, and so provides direct evidence of flow inside the Earth.

Over the past decade, various studies

have revealed strong anisotropy in the Earth's upper mantle, and in the bottom 200–300 km of the lower mantle. But most of the lower mantle has appeared to be largely isotropic to seismic waves, and only a few investigations have hinted at anisotropy in the transition zone between the upper and lower mantle at 410–660 km. On page 777 of this issue, however, Wookey *et al.*¹ report evidence for significant anisotropy at around 660 km, in the mid-mantle between the Tonga subduction zone and Australia. They propose that the anomaly is related to downwelling material

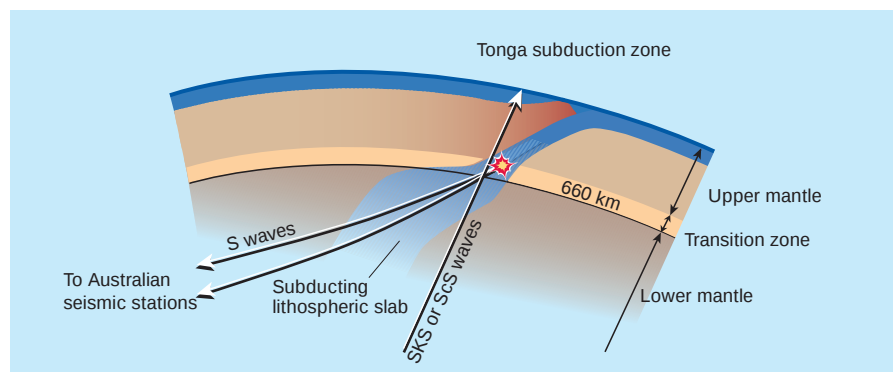


Figure 1 Evidence for anisotropy at depths of around 660 km in the Earth, the boundary between the upper and lower mantle. Wookey *et al.*¹ have analysed S waves produced by deep earthquakes in the Tonga subduction zone, where a slab of lithosphere (crust and uppermost mantle, shown in blue) is diving into the deeper mantle. The seismic stations are in Australia, and the S-wave splitting identified by Wookey *et al.* indicates anisotropy that is consistent with deformation of the subducting lithosphere as it enters the lower mantle; that is, shearing in and close to the slab may create a fabric of aligned minerals in the lower mantle. SKS and ScS are seismic waves that propagate on almost vertical paths at these depths. In this region they are not split from depths greater than 410 km, placing constraints on the geometry of the anisotropy near 660 km.

that is being distorted as it enters the lower mantle.

The velocities of seismic waves are determined by the strength and density of rock through which the waves are passing. Those properties, in turn, are a function of temperature, pressure, the mineral constitution of the rock, and the phase the minerals are in.

Anisotropy in seismic wave velocities arises in two different ways. First, many minerals in the mantle have an intrinsically anisotropic crystal structure. When such crystals are aligned over mantle regions that are large enough to be sensed by seismic waves, variations in wave velocity occur both with propagation direction and with polarization of particle motion. For example, much of the anisotropy in the upper mantle can be attributed to a mix of anisotropic olivine and orthopyroxene crystals that have been aligned either by present-day mantle flow or, in some regions of the lithosphere (the crust and uppermost mantle), by past deformation. Second, anisotropy in wave velocities can be caused by mineral fabrics such as alternating layers with different chemical compositions, or physical phases, when the scale of the fabric is small compared with the wavelength of a seismic wave.

When a seismic wave known as a shear wave enters anisotropic media, its particle motion is 'split' into two orthogonal components with different propagation velocities. The fast and slow vibration directions reflect the intrinsic geometry of the anisotropy, and the time delay between them is a function of the strength of the anisotropy and the propagation distance through the anisotropic region. The shear waves studied by seismologists are classified according to the routes they have taken: S waves have passed directly through the mantle, whereas SKS waves have passed through the core as well, and ScS waves have reflected from the top of the core. Each type of wave conveys different information.

Wookey *et al.*¹ measured splitting in S waves that propagated from earthquakes in the Tonga subduction zone to seismic stations in Australia (see Fig. 1 here, and the map on page 778). In the Tonga subduction zone, one lithospheric plate is diving beneath another, deep into the mantle, generating earthquakes to depths of 660 km. The key to Wookey and colleagues' conclusions is that 15 of their 35 highest-quality S-wave splitting times are between 2 and 7 seconds, values that are much larger than those commonly observed.

To isolate the effects of anisotropy in the transition zone and lower mantle, Wookey *et al.* had to quantify wave splitting caused by anisotropy in the upper mantle. They tested a model for the Australian upper mantle that is 'transversely isotropic' (horizontal plane fast, vertical axis slow), and that matches both the lack of strong splitting in SKS

phases that travelled to the Australian stations on nearly vertical paths and the large effects of anisotropy in horizontally propagating surface waves². This model predicts less than 2 seconds of splitting for the S waves from Tonga. Based on the predictions of models that include anisotropy over a variety of depth ranges, the authors conclude that there must be significant anisotropy at depths of 660–900 km.

However, other estimates of upper-mantle anisotropy beneath Australia are potentially in conflict with those used by Wookey *et al.* For instance, different surface-wave investigations produced models of anisotropy with a more complex form^{3,4} and greater depth⁴. But these models^{3,4} are not obviously consistent with the observed absence of strong SKS-wave splitting. And they cannot explain more than 3–4 seconds of the S-wave splitting seen by Wookey *et al.*

Another question concerns the large scatter in the S-wave splitting times observed by Wookey *et al.*¹. In particular, sources in a narrow latitude range (17.5° S to 22.5° S) yielded both the seven largest splitting times and numerous times in the range 0.5–2.0 seconds. In some cases, S phases on similar paths, but with different particle-motion polarizations, produced both large and small splitting times. These measurements require strong lateral variation in the strength of anisotropy, and an anisotropic symmetry that differs from the simple transverse isotropy assumed in the models tested by Wookey *et al.* But given the large splitting times from some of the deepest sources (550 km or more), and the maximum apparent magnitude of upper-mantle anisotropy beneath Australia, it seems that there must be significant mid-mantle anisotropy in this region.

As Wookey *et al.* acknowledge, there are various explanations for the origin of the anisotropy. In Fig. 3 of their paper (page 779) they present models in which the anisotropy is produced by the subducting Tonga lithospheric slab interacting with the top of the lower mantle. One appealing idea is that the anisotropy stems from lower-mantle minerals that have been aligned by dislocation creep in a region around the slab. Dislocation creep is a process of plastic deformation that aligns crystal axes with respect to strain geometry. When deformation occurs by a different process — diffusion creep — no such alignment occurs.

This hypothesis is in striking agreement with the numerical models of McNamara *et al.*⁵ that predict the existence of dislocation creep near slabs and diffusion creep elsewhere in the lower mantle. Shearing of perovskite, the most abundant and anisotropic mineral at the top of the lower mantle, would produce anisotropy that is inconsistent with the observed fast, horizontal S polarizations if the phases propagated parallel to the shear



100 YEARS AGO

Messrs. Sanders and Crowhurst have sent us for examination a number of brilliant lantern slides of birds and other zoological subjects. Photography has been a helpful handmaid to many branches of science, but none of its performances are more widely appreciated than those in the field of natural history. Drawings of animals may have artistic merit, but they do not inspire the feeling of life which is conveyed by good photographs of objects in their natural surroundings. The lantern slides sent by Messrs. Sanders and Crowhurst are from photographs of birds, nests, eggs and young and other living animals taken by Mr. Oliver G. Pike. To lecturers on natural history such true pictures of living creatures must be invaluable, and no better source of encouragement to study nature could be desired. By the side of such beautiful photographic pictures as are now available for projection upon a screen or for the illustration of books, the drawings which did duty in natural history instruction seem but a vain show. Messrs. Sanders and Crowhurst send us with their slides an ingenious arrangement for viewing lantern slides under a low magnifying power. The arrangement, though simple, is very effective, and a pleasant half hour can be passed by using it to look at lantern slides. From *Nature* 13 February 1902.

50 YEARS AGO

There are so many examples of the adaptation of an animal to its environment which at first sight would appear to find their simplest explanation in the supposition that the effects of the environment have become inherited, that theories of this kind have continued to retain a following in spite of the lack of clear experimental evidence in their support. This following has been composed mainly of naturalists; experimentalists and geneticists have recently tended to adopt an attitude similar to that of Dobzhansky, who writes: "This question has been discussed almost *ad nauseam* in the old biological literature... so that we may refrain from the discussion of it altogether". In dismissing the matter so cavalierly, Dobzhansky was explicitly referring to "direct adaptation", that is, the hypothesis that when the environment produces an alteration in the development of an animal, it simultaneously causes a change in its hereditary qualities such that the developmental alteration tends to be inherited.

From *Nature* 16 February 1952.

plane^{6,7}. But if the S paths were significantly inclined with respect to the shear plane, the observations could possibly be matched¹. Roughly horizontal shearing of perovskite associated with a deforming Tonga slab⁸ may therefore explain the S-wave observations of Wookey *et al.*¹ as well as the apparent absence of large splitting from the lower mantle in SKS and ScS phases that travel almost vertically beneath the Tonga subduction zone^{9,10}.

The work of Wookey *et al.*¹ and McNamara *et al.*⁵ suggests that strong anisotropy may be a common feature of regions where subducting slabs enter the lower mantle. Accurately modelling the full waveform interaction of seismic phases with anisotropic structure, rather than just their arrival times, is a formidable task. But applying such approaches to waves that sample the mid-mantle could help to explain the dynamics of mantle convection. ■

Karen M. Fischer is in the Department of Geological Sciences, Brown University, Providence, Rhode Island 02912, USA.

e-mail: karen_fischer@brown.edu

1. Wookey, J., Kendall, J.-M. & Barruol, G. *Nature* **415**, 777–780 (2002).
2. Gaherty, J. B. & Jordan, T. H. *Science* **268**, 1468–1471 (1995).
3. Debayle, E. & Kennett, B. L. N. *Earth Planet. Sci. Lett.* **184**, 339–351 (2000).
4. Simons, F. J., van der Hilst, R. D., Montagner, J.-P. & Zielhuis, A. *Geophys. J. Int.* (in the press).
5. McNamara, A. K., Karato, S. & van Keken, P. E. *Earth Planet. Sci. Lett.* **191**, 85–99 (2001).
6. Karato, S. *Pure Appl. Geophys.* **151**, 565–587 (1998).
7. Stixrude, L. in *The Core–Mantle Boundary Region* (eds Gurnis, M., Wyssession, M., Knittle, E. & Buffett, B.) 83–96 (American Geophysical Union, Washington, DC, 1998).
8. Van der Hilst, R. *Nature* **374**, 154–157 (1995).
9. Fischer, K. M. & Wiens, D. A. *Earth Planet. Sci. Lett.* **142**, 253–260 (1996).
10. Iidaka, T. & Niu, F. *Geophys. Res. Lett.* **28**, 863–866 (2001).

Developmental biology

Precision patterning

Nipam H. Patel and Sabbi Lal

Many developmental events depend on gradients of key molecules to tell cells where they are. These gradients vary between embryos, but a noise-filtering mechanism ensures that development proceeds normally.

Embryos develop in a precise pattern that varies little from individual to individual within a given species. Such precision and reproducibility make this an excellent example of biological robustness. For instance, embryos tend to be patterned normally, regardless (within reason) of how large or small they are, and even in the face of environmental perturbations such as fluctuations in temperature. This seems remarkable given that certain patterning steps rely on molecular gradients to tell cells where they are in the embryo, and hence how they should develop. Variations in temperature and size would be expected to disrupt such gradients and to lead to errors later on in development.

On page 798 of this issue, Houchmandzadeh and colleagues¹ investigate this robustness by taking a close look at one of the classic gradient systems — that of the Bicoid protein — in embryos of the fruitfly *Drosophila melanogaster*. They find that the gradient can vary greatly from embryo to embryo, but that the positional readout is still quite precise. In other words, the noise in the system is filtered out.

Proteins such as Bicoid are known as morphogens, and work in the following way.

First, a localized source of morphogen is established in the embryo. Subsequent diffusion (or active movement) and degradation of the morphogen produces a concentration gradient that imparts information to cells about their position within a developmental field. So, for example, the gradient of Bicoid protein in fruitflies tells cells where they are along the head-to-tail (anterior–posterior) axis of the embryo. Bicoid is provided in the form of messenger RNA (mRNA) from the mother during egg development and becomes localized to the anterior pole. As the mRNA is translated after fertilization, the resulting protein begins to diffuse from its source. It is not hindered by cell boundaries because the embryo initially develops as a syncytium, in which thousands of cell nuclei share the same cytoplasm.

The formation of the Bicoid protein gradient is presumably influenced by such factors as the quantity of maternal mRNA, the precision of mRNA localization, the rate of translation into protein, and the activity of enzymes that subsequently degrade the protein. All of these processes may be susceptible to perturbations. So how much does the Bicoid gradient differ from embryo to embryo? Houchmandzadeh *et al.*¹ carefully measured the profile of Bicoid protein in a large number of wild-type embryos, and found that it is indeed quite variable (Fig. 1). After normalizing the profile for embryo length, the position at which the Bicoid concentration crosses a chosen level can be anywhere within a region encompassing about 30% of the length of the embryo.

One of the functions of Bicoid protein is to control the production of mRNA (transcription) from a series of ‘gap genes’ in different spatial domains along the embryo’s anterior–posterior axis. One of these, the Hunchback gene, is transcribed in an anterior domain, with a sharp boundary of expression about halfway down the embryo. Previous studies^{2,3} suggested that the Hunchback gene is switched on in response to a specific threshold concentration of Bicoid protein. Indeed, embryos produced by mothers containing two or four extra copies of the Bicoid gene (introduced as ‘transgenes’) show a prominent posterior shift in the position of the Hunchback boundary³. So, the posterior edge of Hunchback expression apparently acts as a direct readout of the concentration of Bicoid protein.

Yet when they looked again at wild-type

Animal behaviour

Snap judgements

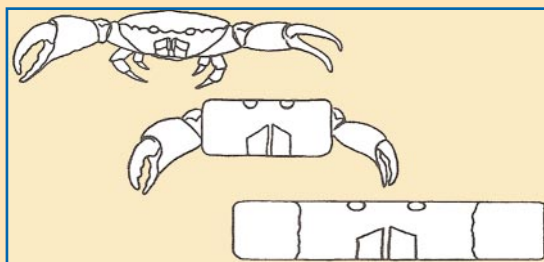
Twice a day, coinciding with tidal motion, climbing crabs *Sesarma leptosoma* leave the canopy of the mangroves they inhabit and migrate down the trunks to the swamp below. But nature being what it is, they run the risk of being eaten on their journey by another crab, *Epixanthus dentatus*.

Stefano Cannicci *et al.* decided to investigate what visual cues alert the climbing crabs to the predator (*Animal Behaviour* **63**, 77–83; 2002). They wrapped two 70-cm lengths of mangrove trunk with PVC sheeting, and placed raffia on the top to create a test pathway. The authors then recorded their subjects’ response to the dummies shown here.

The first is a preserved specimen of *E. dentatus* in ambush posture. The second is a wooden rectangle, painted naturalistically, but with real claws attached. The third is a wooden trapezium, the same size and coloration as *E. dentatus*, but otherwise not like the real thing. Cannicci *et al.* kept all the dummies in mud for a week to give them an authentic

swampy tang. In tests, *S. leptosoma* recognized the first two dummies as dangerous from about 30 cm away, and took avoiding behaviour, but they were unperturbed by the third. So it seems that claws in particular provide the warning. With up to six predators per tree, the climbing crabs probably don’t get a second chance to mistake identity.

Tim Lincoln



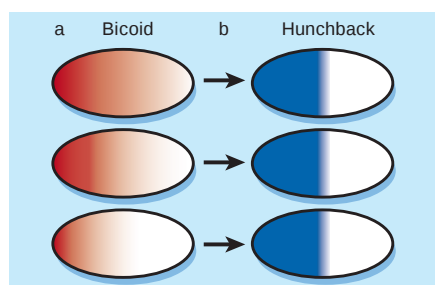


Figure 1 Filtering out noise in development. a, Houchmandzadeh *et al.*¹ find that, within a group of wild-type fruitfly embryos, the profile of the Bicoid gradient (red) can vary significantly (exaggerated somewhat here). b, However, the profile of Hunchback expression (blue) is relatively precise and constant, even though Bicoid regulates Hunchback expression. In fact, the addition of extra copies of the Bicoid gene does lead to shifts in the Hunchback pattern³ (not shown). But Houchmandzadeh *et al.* find that these shifts are not as great as might be expected if the Hunchback pattern depended solely on Bicoid concentration.

embryos, Houchmandzadeh *et al.*¹ found that the Hunchback boundary — at the level of both mRNA and protein — is very precise, showing far less variation than the Bicoid gradient (Fig. 1). Normalizing for egg length, about two-thirds of the embryos showed a precise Hunchback boundary in a range that spanned only about 1% of the total egg length, a precision equivalent to the width of a single nucleus. Furthermore, unlike the Bicoid gradient, the Hunchback boundary position directly correlates with egg length, suggesting that information about proportion is included in the Hunchback expression pattern.

This observation, that the precision of the wild-type Hunchback boundary is unaffected by variations in the Bicoid gradient, seems to be at odds with the finding³ that increasing the number of maternal Bicoid transgenes leads to posterior shifts in the Hunchback boundary. However, when Houchmandzadeh *et al.* repeated the transgene experiments they found that the shift in Hunchback expression was clear, but smaller than expected. When they raised embryos at different temperatures they found that, although the Bicoid protein profile at equivalent developmental stages is significantly altered by temperature changes, the position of the Hunchback boundary is almost unaffected. The implication is that this boundary is subject to correction mechanisms that filter out variability in the Bicoid gradient, as well as a mechanism that imparts scaling information.

Houchmandzadeh *et al.* conclude that the precision of the Hunchback boundary is independent of Bicoid. So what does regulate the Hunchback boundary? The most obvious candidates are other embryonic gap genes. The authors show, however, that eliminating any one of these genes has little or no effect on the absolute position of the Hunchback boundary

and, more importantly, has no effect on the boundary's precision. The authors screened 80% of the fruitfly genome by eliminating entire chromosomes, and still found no embryonically expressed gene that disrupts the precision of the Hunchback boundary.

The most obvious maternally derived candidates — the posteriorly focused gradient of Nanos protein, and maternal Hunchback — affect the position of the embryonic Hunchback boundary but not its precision¹. But Houchmandzadeh *et al.* identified specific mutant forms of the maternal *Staufen* gene that did affect this precision. *Staufen* is known to affect the localization of maternally derived mRNAs at both embryonic poles⁴, but Houchmandzadeh *et al.* show that the effect of *Staufen* on the Hunchback boundary appears to be independent of an effect on Bicoid localization.

These results have several implications. First, the obvious embryonic candidates for regulating the Hunchback boundary are not solely responsible for its precision. So theories proposing that interactions between gap genes are responsible for generating precise expression boundaries may be missing a crucial component. The identification of mutant forms of the *Staufen* gene that affect precision might provide the key to unravelling the mechanism.

Second, it is thought that the Bicoid gene evolved relatively recently, within the Dipterans (the large group of insects that includes *Drosophila*). But evolutionarily distant insects such as grasshoppers also show anterior domains of Hunchback expression⁵ that are presumably wholly independent of Bicoid. So maybe the genetic system that produces precise Hunchback boundaries in *Drosophila* will give us clues to the regulation of Hunchback in these other insects.

Finally, the phenomenon of noise filtration may be a general property of morphogenetic systems, made necessary by their inherent susceptibility to perturbations. Indeed, studies⁶ of the morphogen Dpp in the *Drosophila* wing disc indicate that the concentration gradient of Dpp can differ from its 'activity' gradient (its output). So mechanisms that correct and modulate morphogen gradients may be common to, and required for, the precise, robust and elaborate patterning of different developmental fields. ■

Nipam H. Patel and Sabbi Lal are in the Department of Organismal Biology and Anatomy, and the Howard Hughes Medical Institute, University of Chicago, 5841 South Maryland Avenue, MC1028, Chicago, Illinois 60637, USA. e-mails: npatel@midway.uchicago.edu slall@midway.uchicago.edu

1. Houchmandzadeh, B., Weischaus, E. & Leibler, S. *Nature* **415**, 798–802 (2002).
2. Driever, W. & Nusslein-Volhard, C. *Nature* **337**, 138–143 (1989).
3. Struhl, G., Struhl, K. & Macdonald, P. M. *Cell* **57**, 1259–1273 (1989).
4. Micklem, D. R. *Dev. Biol.* **172**, 377–395 (1995).
5. Patel, N. H. *et al. Development* **128**, 3459–3472 (2001).
6. Teleman, A. A. & Cohen, S. M. *Cell* **103**, 971–980 (2000).

Daedalus

Fresh flavours

Our past as hunters and gatherers has left us with distinctive taste in food. We like it fresh. An animal or vegetable that was living and growing only a few minutes ago has quite a different taste to one that has been stored. Many foods, from sprouts to fish, lose their pleasant flavour very quickly. Even the British diet would be delightful if it were fresh. But in bulk farming, a large amount of food is harvested at the same time and is then stored. This is clearly at variance with our animal nature. But this type of farming is highly efficient, and so has sadly become a fact of life.

While an animal or vegetable is alive, its immune system protects its evanescent compounds or regenerates them. When it dies, all this stops. Bacterial attack, crosslinking and decomposition all start at once. Freezing, that brutal attempt to stop the clock, seems to work best with the bulk components. One food-processing company claims to freeze its vegetable product within 2 hours of picking it, hoping to trap the brief trace compounds of freshness while they last. Daedalus also recalls how the makers of instant coffee put a key flavour volatile in the space at the top of each jar, so the illusion of the real thing survives at least for a moment. DREADCO biochemists are now studying the trace compounds present in fresh foodstuffs.

This delicate and tricky work must be done quickly, using food picked or killed and transported to the lab with equal rapidity. For each foodstuff, Daedalus hopes to identify or synthesize just those elusive volatiles that restore the illusion of freshness to the long-stored product. Farming, that dreary but efficient business, will at last be matched to our instinctive nature.

Standard condiments, such as salt, pepper or monosodium glutamate, are 'amplifiers': they exaggerate whatever taste the food has at the time. By contrast, each DREADCO 'elixir of freshness' will restore the food's own character, so it tastes fresh again. Like pepper, it will be added at the table rather than in the pot. It may take the form of an inert tasteless powder with an added volatile, or a spray-can of liquid or vapour. Daedalus cannot guess how many will be needed. In the worst case, every foodstuff will need its own elixir. But with luck, only a few elixirs will be needed to revive that elusive sense of freshness — one for meat, say, and one for vegetables. Even calorie-counting, vegetarianism and other dietary extremes will gain new pleasure and respectability.

David Jones

Stem cells that make stems

Detlef Weigel and Gerd Jürgens

Plant stem cells, contained in specialized structures called meristems, have amazing regenerative powers. They enable plants to grow and produce new organs throughout lifetimes that can span hundreds of years.

Few animals reach their hundredth birthday. But, although plants have not yet discovered the secret of immortality, many trees live for hundreds of years, and some — such as the bristlecone pines found in the mountains of western North America — are several thousand years old. Plants can achieve these amazing lifespans because their body plan is modular, characterized by continuous and repetitive formation of new structures and organs such as leaves and flowers, which enables them to survive substantial damage.

A prerequisite for plants' longevity is their ability to maintain populations of unspecialized immature cells in the adult plant over very long periods of time. These are the 'stem cells', which can both renew themselves and give rise to many mature specialized cell types. Notwithstanding an unfortunate quirk of the English language, plant stem cells give rise to far more than stems — they are the cells from which all plant tissues derive.

Stem cells in general hardly need any introduction, as this term has become ubiquitous even in newspapers. The reason, of course, is the unsuspected versatility of mammalian stem cells and the recent success in isolating human stem cells, which raises the possibility of using them to treat many human diseases¹.

Stem cells that can produce the entire spectrum of cell types found in an organism are 'totipotent' (from the Latin *totus*, entire). In most animals, the only cell that is truly totipotent is the fertilized egg — the zygote and its immediate descendants. Stem cells in an adult mammal, such as those that continually generate blood cells, give rise to a restricted spectrum of mature cell types and are considered to be merely pluripotent (from the Latin *pluris*, many or several). Many plant cells, in contrast, continue to be totipotent throughout the plant's life — plants can be propagated from small pieces of tissue or even from single cells.

Until recently, however, we have known little about the molecules that control the fate of plant stem cells or that organize them within the meristems, the specialized structures present at the tips of shoots and roots. But this is now changing. Much recent progress in understanding the function of meristems and plant stem cells has come from studies in the thale cress *Arabidopsis*

thaliana, a small weed popular for research and the first plant to have its genome completely sequenced. Many of the principles of stem-cell regulation elucidated in *Arabidopsis* will probably also apply to other plants, as mutations in the corresponding genes in snapdragon (*Antirrhinum*) or maize (*Zea mays*) cause similar defects. Here we focus on the most salient features of meristem development and organization (see ref. 2 for a more detailed review), and discuss how interactions between the different kinds of cells in a meristem keep it functioning properly.

Embryonic origin of meristems

A plant seedling typically starts out with two meristems — situated at the tips of the root and shoot respectively. These are the primary meristems, which produce most of the adult plant (Fig. 1). The primary root meristem produces all the tissues of the main root, and the primary shoot meristem all the tissues of the main plant stem, including the leaves. The latter also gives rise directly to secondary meristems that produce side shoots and flowers.

Both primary meristems are formed early in embryonic development, as illustrated in Box 1 (overleaf). To establish the root meristem one cell is singled out to become the hypophysis or precursor of the so-called quiescent centre, which organizes the meristem. The establishment of the quiescent centre itself seems to depend on cell–cell communication mediated by the plant hormone auxin, a small molecule chemically related to the amino acid tryptophan. That auxin is involved has been inferred from the observation that embryos lacking the auxin-response factor MONOPTEROS (MP) fail to single out the hypophysis, and no root tissue develops³.

A similar inductive process involving cell–cell communication probably defines the shoot meristem. One of the recent advances in understanding meristem development has been the identification of molecular markers, which are particular genes (and the proteins they encode) that are characteristic of the cell, structure or stage in question. The first molecular marker that identifies the developing shoot meristem is the gene *WUSCHEL* (*WUS*), which is initially switched on in four inner cells of the embryo during the 16-cell stage (see Box 1).

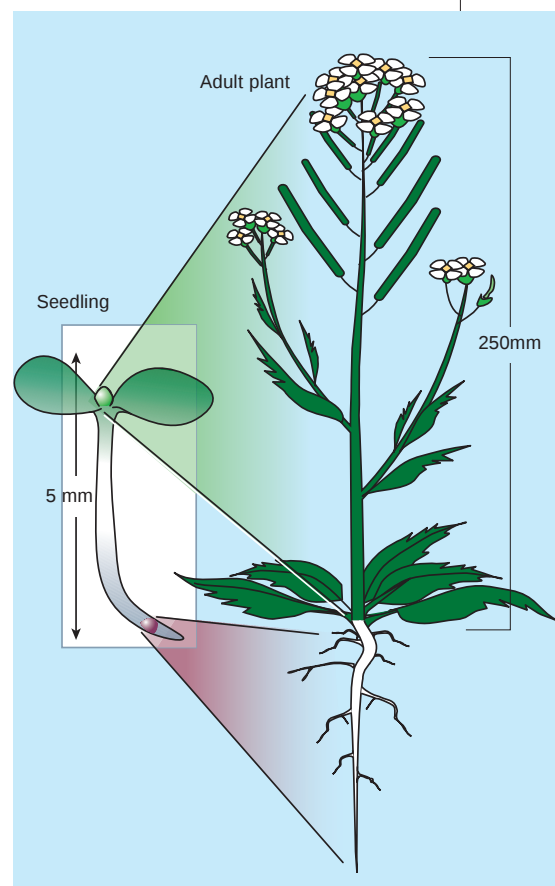


Figure 1 **Contribution of shoot and root meristems to the *Arabidopsis* plant.** The shoot meristem contains some 100 cells and is less than 100 μm in diameter. The root meristem is similar in size. Approximate scale bars are shown for comparison.

When these cells divide, only the cells adjacent to the precursors for the embryonic vascular system maintain *WUS* expression, suggesting that a signal from the vascular precursor cells is required to maintain *WUS* expression.

Meristem organization

Stem cells are confined to the centres of shoot and root meristems. They divide slowly and give rise to two types of daughter cell. The cells that stay in the centre remain stem cells, whereas daughter cells that are displaced to the periphery of the meristem enter a developmental pathway that leads to differentiation into root- or shoot-specific cells or specific organs such as leaves, or to initiation

of secondary meristems (see, for example, refs 4–7). As a meristem often has to function over a long period of time, the balance between stem-cell proliferation and differentiation needs to be extraordinarily precise. Just one extra cell division once a year, for example, would make a meristem 1,000 times larger than normal after just 10 years — a short time in the life of a tree.

Stem cells and their proliferation are maintained by signals that cells receive from the local environment. At the heart of preserving the organization and functioning of the shoot meristem is a negative-feedback loop with two critical elements, the *WUS* and *CLAVATA3 (CLV3)* genes. *WUS* encodes a transcription factor, a protein that directly regulates the activity of other genes. It is expressed in a central population of cells, known as the organizing centre, immediately below the apical layer of active stem cells⁸ (see Box 1). By an unknown mechanism, *WUS* tells adjacent cells to be stem cells, which are in turn characterized by expression of *CLV3*. The secreted *CLV3* protein, in turn, interacts with the more widely expressed *CLV1/CLV2* receptor complex to limit the area of *WUS* expression in the underlying organizing centre^{9–12}.

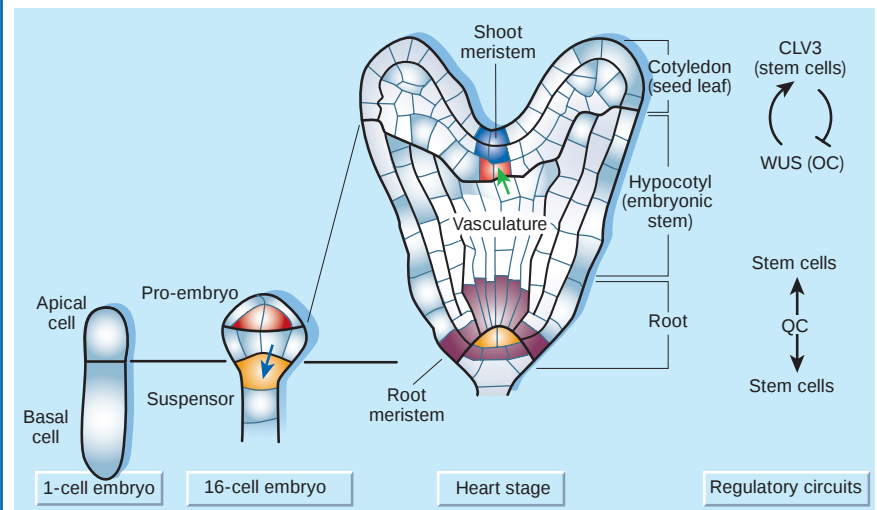
This negative-feedback loop elegantly corrects transient aberrations in stem-cell number. An excess of stem cells leads to an excess of *CLV3*, which causes a reduction in *WUS* expression and consequently a reduced signal for stem-cell proliferation. If, on the other hand, there are too few stem cells, the deficit in *CLV3* leads to an increase in *WUS*, which in turn increases stem-cell number. The importance of this feedback loop is revealed when it is permanently disturbed. For example, if the *WUS* gene is knocked out, or *CLV3* expression is made independent of *WUS*, the meristem shuts down. Conversely, if *CLV3* is knocked out, or *WUS* expression is made independent of the negative-feedback loop, the meristem enlarges dramatically.

Although we do not know whether a comparable regulatory circuit controls the function of the root meristem, its organization is similar to that of the shoot meristem, with an organizing centre of non-dividing cells — the quiescent centre — being essential to maintain the stem-cell fate of the immediately adjacent cells (see Box 1). As in the shoot, cell–cell communication is essential: if a quiescent-centre cell is killed (using a precisely focused laser beam), the adjacent cells will instead differentiate¹³.

From meristems to organs

The *raison d'être* of a meristem is not just to maintain itself, but to produce the various parts of the plant. The primary shoot and root meristems both give rise to elongating structures, the primary stem and the primary root, both of which are decorated

Box 1 Embryonic origin of primary meristems



The single-celled *Arabidopsis* zygote (not shown) elongates in a direction that indicates the future apical–basal orientation of the embryo; a horizontal division produces the one-celled embryo. The apical cell produces most of the embryo, and the basal cell produces a filamentous suspensor six to eight cells long that connects the embryo to the maternal tissue. A stereotyped series of cell divisions produces the 16-cell pro-embryo (of which only 8 cells are shown here). A signal from the pro-embryo (bright blue arrow) is thought to establish the fate of the hypophysis (orange), the uppermost cell of the suspensor, which gives rise to the quiescent organizing centre of the root meristem. The first shoot-meristem marker to be

expressed is *WUS* (red), which can be detected in the 16-cell pro-embryo. Its expression is anchored in the centre of the embryo, so that asymmetric cell divisions give rise to *WUS*-positive and *WUS*-negative cells. As can be seen at the heart stage, the *WUS*-positive cells of the shoot meristem have remained next to the precursors of the vascular system, which suggests that *WUS* expression is induced and/or maintained by a signal from these cells (green arrow). Immediately above them are the stem cells of the shoot meristem (dark blue).

Once the heart-stage embryo, with about 250 cells, has been produced, the entire basic body plan of the embryo has formed. The apical–basal

pattern along the main axis consists of: the shoot meristem flanked by the two regions that will give rise to the cotyledons, or seed leaves; the embryonic stem, or hypocotyl; the root; and the root meristem.

Populations of stem cells within meristems are maintained by adjacent organizing centres. In the shoot meristem, the negative-feedback loop of *WUS/CLV3* (see text) regulates the size of the stem-cell population. *WUS*-expressing cells of the organizing centre (OC; red) maintain the overlying cells as stem cells (dark blue). In the root meristem, the non-dividing cells of the quiescent centre (QC, orange) maintain the adjacent cells as stem cells (purple).

D.W. & G.J.

by lateral organs such as side shoots and lateral roots. Despite this superficial similarity, the ways in which lateral organs are generated in the shoot and root are quite different. Leaves are typically produced directly by the primary shoot meristem, whereas lateral roots are derived from secondary root meristems that initiate *de novo* from an internal tissue layer of the growing root, as we discuss later.

In the shoot meristem, daughter cells displaced to the periphery of the meristem enter a 'zone of competence' where new organs are formed. A new organ such as a leaf starts out as a group of still undifferentiated cells, which respond coordinately to a (still unknown) signal to form the incipient organ, called a primordium. In *Arabidopsis*, leaf primordia derive from groups of about a dozen 'primordium founder cells' that are

contributed by both the epidermal and subepidermal layers of the meristem periphery^{14,15}. Two factors limit the area from which an individual primordium will form: distance from the centre of the meristem, where the stem cells reside, and distance from previously formed primordia. Cells that end up between primordia eventually leave the zone of competence and contribute to the stem. We do not fully understand the signals that position primordia, but it seems that auxin is important here as well^{16,17}.

Several markers distinguish the proliferation zone of a meristem from the incipient primordia, including the *SHOOT MERISTEMLESS (STM)* gene, which prevents cells from entering a differentiation pathway. It does this, at least in part, by repressing another gene, *ASYMMETRIC LEAVES1*

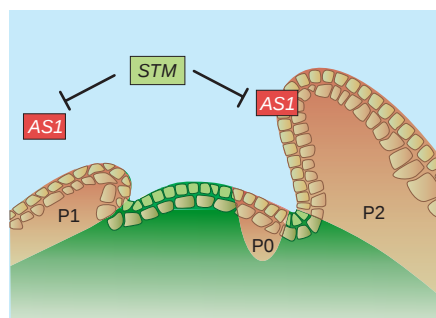


Figure 2 The primary shoot meristem of a vegetatively growing plant. The shoot meristem in the centre is flanked by two leaf primordia. The *STM* gene (bright green) maintains the undifferentiated state by repressing *AS1* (red), which is expressed in primordium founder cells (P0) as well as young primordia (P1, P2). Note the zonation of the meristem, with clearly demarcated epidermal and subepidermal cell layers. The cells underneath the subepidermal cell layer are not shown.

(*AS1*), whose expression is normally restricted to primordium founder cells¹⁸ (Fig. 2). In mutants lacking *STM*, *AS1* expression expands into the meristem, causing the meristem to shut down. Conversely, expression of *AS1* represses other meristem regulators such as the *STM*-related gene *KNAT1* in the emerging primordia. In mutants lacking *AS1*, expression of *KNAT1* persists in the leaf primordia, resulting in 'leaves' with shoot-like characteristics^{18,19}. Like *WUS*, the differentiation regulators *STM*, *AS1* and *KNAT1* encode transcription factors.

The primary meristems share the work of producing the often enormously complex shoot and root systems of adult plants with several kinds of secondary meristems. The organization of a secondary meristem is essentially identical to that of the primary meristems. Two examples of secondary meristems in the aerial part of herbaceous plants such as *Arabidopsis* are secondary shoot meristems, which are positioned in the junctions of leaves with the stem, and floral meristems, which produce the flowers. The floral meristems, at least, are produced directly from the primary shoot meristem²⁰ (Fig. 3). Although the shoot and floral meristems are contiguous, the expression of stem-cell regulators such as *WUS* and *CLV3* in the centre of the meristems is not, so their expression must be re-established *de novo* in the floral meristem^{8,9}. The plant must, therefore, have a process of self-organization that generates a secondary meristem — a process that may be analogous to the initiation of primary meristems during embryo development.

An important difference between floral and shoot meristems, both primary and secondary, is that a flower is typically composed of a fixed number of parts (in *Arabidopsis*, four sepals, four petals, six

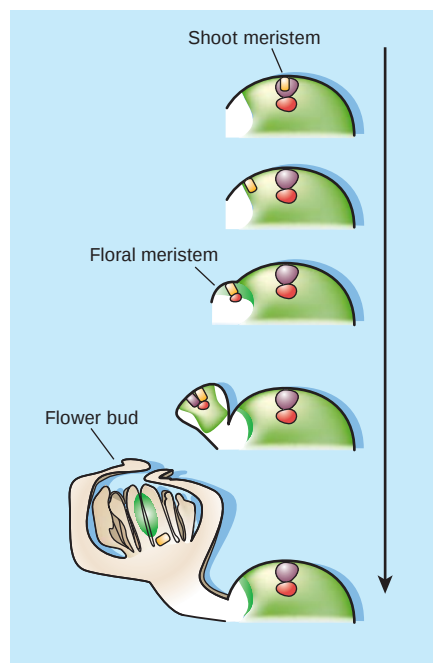
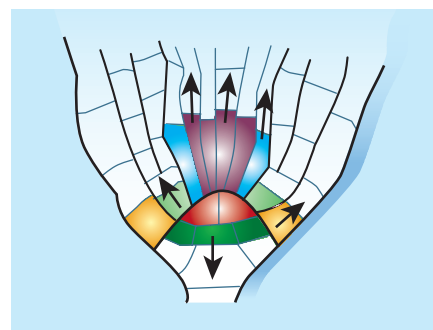


Figure 3 The developmental fate of a cell is determined by its position. A derivative of a stem cell (yellow) is displaced to the periphery of the shoot meristem, becomes part of a floral meristem and is incorporated into a flower (divisions of this cell that take place during this developmental journey have been ignored for simplicity). Expression of the meristem marker *STM* is shown in bright green, expression of the stem-cell marker *CLV3* in purple, and expression of *WUS* in red. Note how this cell loses expression of *CLV3* as it leaves the shoot meristem, but reacquires it in the floral meristem. This cell loses its stem-cell identity again as it is displaced from the proliferating centre (indicated in green) of the developing flower.

stamens and two carpels), whereas a shoot meristem can potentially generate an unlimited number of leaves. The limited function of the floral meristem is reflected in a different pattern of *WUS* expression, which is only transiently maintained in flowers, whereas it persists indefinitely in shoot meristems. If *WUS* is not shut down in the flower, as happens in plants that have a mutation in the *AGAMOUS* (*AG*) gene, the floral meristem behaves like a shoot meristem and continues to produce an



indefinite number of parts. Notably, *WUS* itself is important in inducing expression of its negative regulator *AG*, which has a general role in specifying the centre of the flower^{21,22}. In contrast to the other *WUS* target, *CLV3*, continued expression of *AG* does not require *WUS*, thus allowing *AG* expression to persist even after *WUS* has been switched off.

The primary root meristem has a somewhat simpler task compared with the shoot meristem, as it gives rise only to the tissues of the main root; lateral roots arise from secondary meristems initiated *de novo*, as we will see later. The primary root meristem consists of two tiers of stem cells immediately above and below the non-dividing quiescent centre. The upper-tier stem cells produce most of the root tissues, with each stem cell giving rise to tissue-specific cell types (Fig. 4). Although this stereotypic mode of producing different cell types suggests a lineage-dependent process of cell-fate determination, the fate of the differentiating cells is not in fact imposed by the stem cells from which they derive, but rather by signals from differentiated cells within the same tissue layer²³. The daughter cells of the lower-tier stem cells contribute to the root cap, which shields the root tip from destruction as it moves through the soil and has to be continuously replaced as the older cap cells wear off.

Unlike the floral meristems, the secondary meristems that produce the lateral roots are initiated *de novo* by small groups of cells within an internal tissue layer of the root called the pericycle. After receiving a signal, which involves auxin, mature cells in the pericycle lose their differentiated status and start to divide. Interestingly, a self-maintaining lateral root meristem forms only after these newly generated cells have begun to produce differentiated tissues typical of a root²⁴.

In the aerial parts of plants there are also secondary meristems that are not contiguous with the primary meristems and that arise *de novo*. Examples are intercalary meristems of grasses, which contribute to the lengthening of the stem, and the cambial meristems of woody plants, which are responsible for increasing the girth of the stem. Even shoot meristems can form *de*

Figure 4 The primary root meristem in the embryo. The black arrows indicate how stem-cell division produces regular vertical files of cells. Above the quiescent centre (red) are the stem cells for the root tissues: lateral root cap and epidermis (yellow), cortex and endodermis (light green), pericycle (light blue) and vascular tissue (purple). Below are the stem cells for the central portion of the root cap (dark green). The meristem continues to work in the same way in the root after germination.

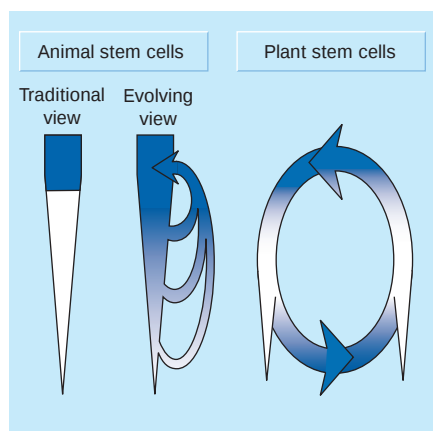


Figure 5 Stem-cell concepts. The fate change of a single cell is shown, with blue indicating stem-cell properties, and white differentiation. The potential to act as a stem cell is indicated by the intensity of the blue colour. In the traditional view, animal stem cells underwent an abrupt transition from being totipotent or pluripotent to having permanently entered a differentiation pathway. In the evolving view, cells gradually lose their ability to revert to stem-cell fate. In plants, on the other hand, it is not uncommon for differentiating cells to become stem cells again. The left and middle diagrams have been adapted from ref. 28.

novo, although this is unusual. They arise in the context of a normal plant as adventitious shoot meristems, which, for example, produce the shoots that form on tree stumps, or when a plant is regenerated from a differentiated organ part, such as a piece of leaf.

Stem cells in plants and animals

What, then, can we learn from plant stem-cell biology? A characteristic of plant stem cells is that they are strictly defined by their position within the meristem, and one does not even need molecular markers to identify, for example, the stem cells in the root meristem (see Fig. 4). Although this is also true for some animal stem cells, such as those in the gonads of fruitflies or worms²⁵, it seems to be the exception rather than the rule, and molecular markers are normally required to distinguish animal stem cells within a population of morphologically identical cells. Nevertheless, the local environment is also critical for the maintenance of animal stem cells, an observation that has resulted in the concept of stem-cell niches^{25,26}. Some of the molecules required for stem-cell survival, such as *PIWI* in flies and *ZWILLE/PINHEAD* in plants, are closely related, which is surprising given that multicellularity probably evolved independently in plants and animals²⁷.

One of the reasons for the recent excitement over animal stem cells is that they have been found to be much more versatile than anticipated, with the possibility of even apparently differentiated cells reverting to a stem-cell fate²⁸. The traditional view of animal stem cells has been that of a one-way road, from cells that can give rise to the entire spectrum of mature cell types to cells with more and more restricted differentiation potential (Fig. 5). In contrast, the fact that plant cells are extraordinarily versatile has never been questioned. As an example, some of the most important animal stem cells are those that give rise to the germ line, and these are set aside early in embryo development. Plants, on the other hand, do not have a dedicated germ line, but generate germ cells repeatedly during the indepen-

dent formation of each flower from a group of undifferentiated cells.

Given this versatility, is it possible that fully differentiated plant cells become stem cells again? In animals, a standard assay for stem-cell properties is to transplant potential stem cells and to monitor the cell lineages to which they can contribute. Unfortunately, because plant cells are encased in rigid cell walls, one cannot easily transplant single cells in plants. But it is possible to make plant cells undergo an alternative form of embryo formation called somatic embryogenesis, a process that does not require formation of a zygote by fertilization. Somatic embryogenesis can be induced after plant cells have been dissociated and taken into suspension culture; early embryos are usually the best starting material, but adult tissue can also be used. Mutations that increase the size of the embryonic shoot meristem also increase the propensity with which somatic embryos can be induced, thus linking stem-cell function and somatic embryogenesis²⁹.

Somatic embryos can also form on leaves in plants that misexpress in leaf cells either of two transcription factors normally involved in embryo maturation^{30,31}. This observation strongly suggests that even fully differentiated cells can become stem cells again, although it has been argued that these transcription factors simply confer a prolonged 'embryogenic state' on somatic cells³². Nevertheless, all these results clearly show that, in contrast to animals, totipotency in plants is not restricted to the zygote or the very early embryonic cells that arise from it.

Somatic embryogenesis is not the only case in which mature or maturing cells become stem cells again. One example from normal plant development is lateral root development, which is initiated by mature pericycle cells²⁴. These cells truly dedifferentiate, as a marker for cell polarity becomes relocalized from its original position at the basal end of the cells to the outer side, which becomes the new basal end in the lateral root primordium³³. Similarly, during induction of adventitious shoot meristems in leaf explants, *STM* expression (and presumably *CLV3* expression as well)

appears *de novo*³⁴. In contrast to what we have learned about the embryonic origin and organization of primary meristems, we still know much less about the molecular basis for the versatility of differentiated plant cells. Understanding how the process of differentiation can be reversed in plant cells should be illuminating for stem-cell biology in other systems as well.

Detlef Weigel is in the Plant Biology Laboratory, Salk Institute for Biological Studies, La Jolla, California 92037, USA, and in the Department of Molecular Biology, Max Planck Institute for Developmental Biology, D-72076 Tübingen, Germany.

Gerd Jürgens is at the Centre for Molecular Biology of Plants (ZMBP), University of Tübingen, D-72076 Tübingen, Germany.

e-mails: weigel@weigelworld.org

gerd.juergens@zmbp.uni-tuebingen.de

- McKay, R. *Nature* **406**, 361–364 (2000).
- Jürgens, G. *EMBO J.* **20**, 3609–3616 (2001).
- Hardtke, C. S. & Berleth, T. *EMBO J.* **17**, 1405–1411 (1998).
- Laufs, P., Grandjean, O., Jonak, C., Kieu, K. & Traas, J. *Plant Cell* **10**, 1375–1390 (1998).
- Haecker, A. & Laux, T. *Curr. Opin. Plant Biol.* **4**, 441–446 (2001).
- Clark, S. E. *Curr. Opin. Plant Biol.* **4**, 28–32 (2001).
- Benfey, P. N. & Scheres, B. *Curr. Biol.* **10**, R813–R815 (2000).
- Mayer, K. F. et al. *Cell* **95**, 805–815 (1998).
- Fletcher, J. C., Brand, U., Running, M. P., Simon, R. & Meyerowitz, E. M. *Science* **283**, 1911–1914 (1999).
- Trotochaud, A. E., Jeong, S. & Clark, S. E. *Science* **289**, 613–617 (2000).
- Schoof, H. et al. *Cell* **100**, 635–644 (2000).
- Brand, U., Fletcher, J. C., Hobe, M., Meyerowitz, E. M. & Simon, R. *Science* **289**, 617–619 (2000).
- van den Berg, C., Willemsen, V., Hendriks, G., Weisbeek, P. & Scheres, B. *Nature* **390**, 287–289 (1997).
- Bossinger, G. & Smyth, D. R. *Development* **122**, 1093–1102 (1996).
- Irish, V. F. & Sussex, I. M. *Development* **115**, 745–753 (1992).
- Reinhardt, D., Mandel, T. & Kuhlemeier, C. *Plant Cell* **12**, 507–518 (2000).
- Vernoux, T., Kronenberger, J., Grandjean, O., Laufs, P. & Traas, J. *Development* **127**, 5157–5165 (2000).
- Byrne, M. E. et al. *Nature* **408**, 967–971 (2000).
- Ori, N., Eshed, Y., Chuck, G., Bowman, J. L. & Hake, S. *Development* **127**, 5523–5532 (2000).
- Long, J. & Barton, M. K. *Dev. Biol.* **218**, 341–353 (2000).
- Lohmann, J. U. et al. *Cell* **105**, 793–803 (2001).
- Lenhard, M., Bohnert, A., Jürgens, G. & Laux, T. *Cell* **105**, 805–814 (2001).
- van den Berg, C., Willemsen, V., Hage, W., Weisbeek, P. & Scheres, B. *Nature* **378**, 62–65 (1995).
- Malamy, J. E. & Benfey, P. N. *Development* **124**, 33–44 (1997).
- Watt, F. M. & Hogan, B. L. *Science* **287**, 1427–1430 (2000).
- Whetton, A. D. & Graham, G. J. *Trends Cell Biol.* **9**, 233–238 (1999).
- Benfey, P. N. *Curr. Biol.* **9**, R171–R172 (1999).
- Blau, H. M., Brazelton, T. R. & Weimann, J. M. *Cell* **105**, 829–841 (2001).
- Mordhorst, A. P. et al. *Genetics* **149**, 549–563 (1998).
- Lotan, T. et al. *Cell* **93**, 1195–1205 (1998).
- Stone, S. L. et al. *Proc. Natl Acad. Sci. USA* **98**, 11806–11811 (2001).
- de Vries, S. C. *Trends Plant Sci.* **3**, 451–452 (1998).
- Steinmann, T. et al. *Science* **286**, 316–318 (1999).
- Teo, W. L., Kumar, P., Goh, C. J. & Swarup, S. *Plant Mol. Biol.* **46**, 567–580 (2001).

Related websites

- ♦ www.plantsci.cam.ac.uk/Haseloff/teaching/MCBPart1B/Lecture3.html
- ♦ www.arabidopsis.org
- ♦ www.life.umd.edu/classroom/bsci124/main.html

Rational imitation in preverbal infants

Babies may opt for a simpler way to turn on a light after watching an adult do it.

Here we show that if an adult demonstrates a new way to execute a task to a group of infants aged 14 months, the children will use this action to achieve the same goal only if they consider it to be the most rational alternative. Our results indicate that imitation of goal-directed action by preverbal infants is a selective, interpretative process, rather than a simple re-enactment of the means used by a demonstrator, as was previously thought¹⁻³.

In Meltzoff's seminal study¹, a group of 14-month-old subjects watched a demonstrator illuminate a light-box by leaning forwards and touching its top with her forehead^{1,2}. One week later, two-thirds of them re-enacted this head action to achieve the same outcome, although none of the control group used it spontaneously. This was taken as evidence that infants separate the goal from the means, automatically imitating the means as demonstrated². Such imitative learning is thought to be specific to humans, as primates do not imitate new strategies to achieve goals, relying instead on motor actions already in their repertoire (emulation)³. If this were also the case in infants, they would be expected to touch the box with their hands, rather than imitating the unfamiliar head action. (Meltzoff, however, did not report such hand actions^{1,2}.)

The readiness of infants to re-enact the head action is surprising, given that 1-year-old babies can evaluate the rationality of the means in relation to the goal and the constraints of the situation^{4,5}. When constraints change, these infants are able to work out the most effective action that the demonstrator should use to achieve the goal (the principle of rational action^{6,7}). Infants would therefore be expected to re-enact an action only if it seemed to them to be the most effective means to achieve the goal (see also ref. 8).

So why did Meltzoff's subjects re-enact the head action, when they could just have touched the box with their hands? If infants noticed that the demonstrator declined to use her hands despite the fact that they were free, they may have inferred that the head action must offer some advantage in turning on the light. They therefore used the same action themselves in the same situation.

To test this idea, we replicated Meltzoff's study¹ with one modification: in one condition, the subjects could see that the demonstrator's hands were occupied while she executed the head action (pretending to be cold, she had wrapped a blanket around herself which she held onto with both hands; 'hands occupied', Fig. 1a). After witnessing the same head action when the adult's hands

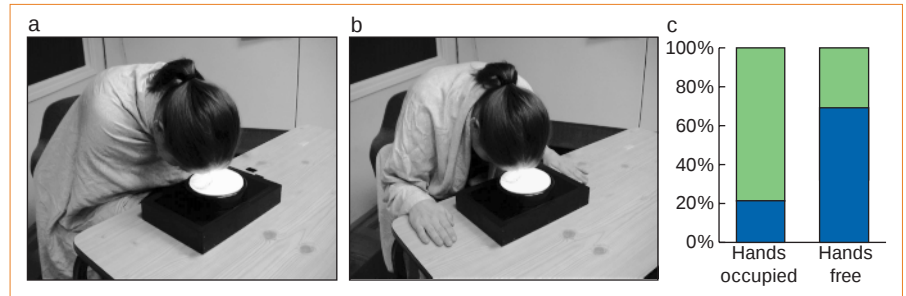


Figure 1 Comparison of the methods used by 14-month-old infants to switch on a light-box 1 week after watching how an adult executed the same task under two different conditions. **a, b**, Adult switching on the light by touching the lamp with her forehead in the hands-occupied condition (**a**, $n = 14$) or the hands-free condition (**b**, $n = 13$). **c**, Methods used by infants to switch on the light-box after watching the head action used by the demonstrator under these two conditions (left bar, adult had hands occupied; right bar, adult had hands free), recorded over a 20-s period. Blue, head action was re-enacted; green, only manual touch was used. Further details are available from the authors.

were free (Fig. 1b), 69% of infants re-enacted the head action, replicating Meltzoff's results¹. However, after watching the adult turn on the light with her head when her hands were occupied, the number of children who imitated the head action dropped significantly to only 21% ($P < 0.02$; Fig. 1c). It must therefore have seemed sensible to the infants that the demonstrator should use the head action when her hands were occupied — nevertheless, 79% of them chose not to imitate her because their own hands were free, presumably concluding that the head action was not the most rational.

Whether they re-enacted the head action or not, all infants who watched the adult perform under both conditions still used the hand action. This suggests that 14-month-old infants are still subject to an automatic, emulation-like process whereby the memory of the effect (illumination by touch) activates the response that is most strongly associated with establishing contact (hand action). But the re-enactment of the head action, when inferred to be rational by the

infant, indicates that imitation by 14-month-olds goes beyond emulation. We conclude that the early imitation of goal-directed actions is a selective, inferential process that involves evaluation of the rationality of the means in relation to the constraints of the situation.

György Gergely*, **Harold Bekkering†‡**,
Ildikó Király*

*Institute for Psychology, Hungarian Academy of Sciences, 1132 Budapest, Hungary
e-mail: gergelyg@mtapi.hu

†Max Planck Institute for Psychological Research, Amalienstrasse 33, 80799 Munich, Germany

‡Present address: Department of Experimental and Work Psychology, University of Groningen, 9712 TS Groningen, The Netherlands

1. Meltzoff, A. N. *Dev. Psychol.* **24**, 470–476 (1988).
2. Meltzoff, A. N. *J. Exp. Child Psychol.* **59**, 497–515 (1995).
3. Tomasello, M. *The Cultural Origins of Human Cognition* (Harvard Univ. Press, Cambridge, Massachusetts, 1999).
4. Gergely, G. *et al. Cognition* **56**, 165–193 (1995).
5. Csibra, G. *et al. Cognition* **72**, 237–267 (1999).
6. Gergely, G. & Csibra, G. *Cognition* **63**, 227–233 (1997).
7. Csibra, G. & Gergely, G. *Dev. Sci.* **1**, 255–259 (1998).
8. Bekkering, H., Wohlschläger, A. & Gattis, M. *Q. J. Exp. Psychol. Human Exp. Psych.* **53A**, 153–164 (2000).

Evolutionary biology

Performance constraints in decathletes

Physical performance by vertebrates is thought to be constrained by trade-offs between antagonistic pairs of ecologically relevant traits and between conflicting specialist and generalist phenotypes^{1,2}, but there is surprisingly little evidence to support this reasoning³⁻⁵. Here we analyse the performance of world-class athletes in standardized decathlon events and find that it is subject to both types of trade-off, after correction has been made

for differences between athletes in general ability across all 10 events. These trade-offs may have imposed important constraints on the evolution of physical performance in humans and other vertebrates.

Decathletes compete in 10 different track and field events over two consecutive days. We constructed a data set of the performance of 600 world-class decathletes across all events from sources available on the Internet (see Fig. 1 legend). Individual performance for any pair of disciplines was positively correlated for the entire data set. This was unexpected, as physiological and biomechanical theory predicts that there should be trade-offs between certain pairs

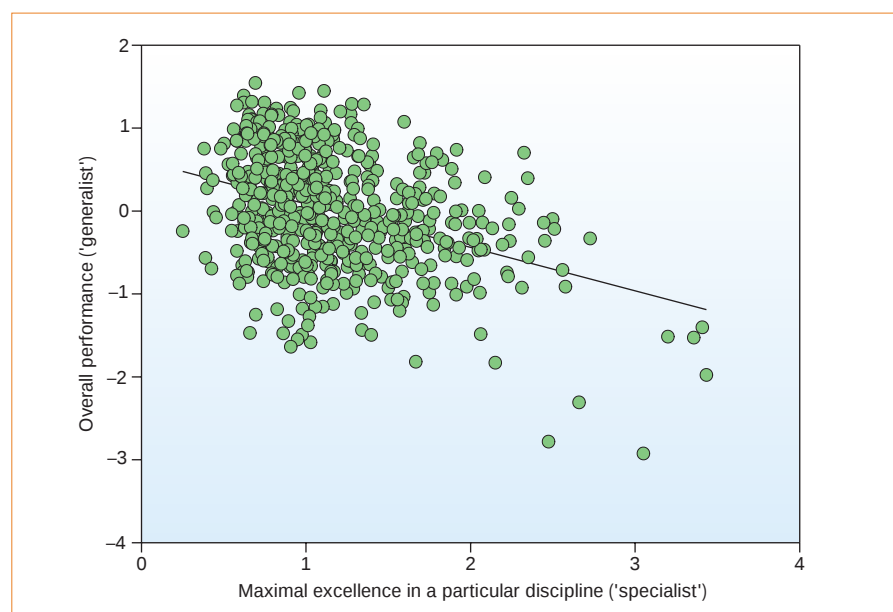


Figure 1 Trade-off between excellence in a single discipline (specialist) and average performance across all 10 events (generalist) for 600 world-class decathletes. Excellence in a particular discipline (represented by the highest residual score for a particular athlete) is negatively correlated with average performance (expressed as the average of all 10 residual scores for that athlete). To ensure that scores in different events received equal weighting, an athlete's score was standardized for each event by subtracting it from the population mean and dividing it by the population standard deviation. The decathlete data set was constructed from the following websites: www.iaaf.org; www.gbrathletics.com; www.athletics.org.au; www.dlv-sport.de; www.athle.org.

of performance traits — for example, speed depends on the athlete having a high proportion of fast, fatigue-sensitive muscle fibres, whereas endurance relies on a higher proportion of slower fibres that are more resistant to fatigue^{3,6–8}.

However, when the analysis was restricted to athletes of comparable ranking (on the basis of International Amateur Athletic Federation scores), trade-offs became evident between certain traits. For example, performance in the 100 metres is negatively correlated with speed over 1,500 metres for athletes who score above 8,000 points ($n = 133$; $r = -0.21$, $P = 0.016$). We suspect that the apparent contradiction in our results is caused by differences in 'general quality' between athletes, which effectively mask any trade-off effects when the entire population is examined.

To account for these differences in general quality, we calculated partial correla-

tion coefficients for each pair of performance traits and found evidence of some performance trade-offs (Table 1). For example, good sprinters performed relatively poorly over 1,500 metres, but did well in the long jump, the 400 metres and the 110-metre hurdles. Likewise, there was a negative correlation between performance in shot-putting (which calls for explosive power in the upper limbs) and in the 1,500 metres (which depends on endurance in the lower limbs).

These results seem to support the idea of trade-offs generated by conflicting anatomical (such as relative body weight and limb proportions) and muscle-fibre-type requirements. Negative correlations may also reflect differences in training, for example if athletes focus on certain disciplines to the detriment of others. However, the consistency with which particular performance traits co-vary suggests that it

is more difficult to combine excellence in some pairs of events than in others.

We calculated measures of excellence and average performance for each decathlete across all disciplines to test the assumption that there must be a performance trade-off between specialist and generalist phenotypes. The principle of allocation, a hypothesis that underlies much of ecological and evolutionary theory, predicts that excellence in one task can only be attained at the expense of average performance in all other tasks, and vice versa¹.

For each athlete, we selected the highest standardized score from his set of 10 events and defined this as that athlete's 'degree of excellence'. We also calculated the average of the 10 scores for each athlete to estimate their overall average performance. The degree of excellence was negatively correlated with average performance across all disciplines ($r = -0.37$, $P < 0.00001$; Fig. 1), which is consistent with the principle of allocation.

We conclude that in an environment in which the selection criterion is combined high performance across multiple tasks, increased performance in one function may impede performance in others. The relevance of our findings for the evolution of performance traits will depend on whether these correlations have a genetic basis — studies of mice⁹ and insects¹⁰ indicate that this may be the case. The influence of an individual athlete's training schedule or career development on these apparent trade-offs also needs to be determined.

Raoul Van Damme, Robbie S. Wilson, Bieke Vanhooydonck, Peter Aerts

Department of Biology, University of Antwerp, Universiteitsplein 1, 2610 Wilrijk, Belgium
e-mail: rwilson@uia.ac.be

- Levins, R. *Evolution in Changing Environments* (Princeton Univ. Press, New Jersey, 1968).
- Futuyma, D. J. & Moreno, G. *Annu. Rev. Ecol. Syst.* **19**, 207–233 (1988).
- Garland, T. Jr in *Quantitative Genetic Studies of Behavioral Evolution* (ed. Boake, C. R. B.) 251–277 (Univ. Chicago Press, Chicago, 1994).
- McPeck, M. A. *Am. Nat.* **148** (suppl.), 124–138 (1996).
- Whitlock, M. C. *Am. Nat.* **148** (suppl.), 65–77 (1996).
- Vanhooydonck, B., Van Damme, R. & Aerts, P. *Evolution* **55**, 1040–1048 (2001).
- Eshbjörnsson, M., Sylvén, C., Holm, I. & Jansson, E. *Int. J. Sports Med.* **14**, 257–263 (1993).
- Rivero, J. L., Serrano, A. L., Henckel, P. & Agüera, E. *J. Appl. Physiol.* **75**, 1758–1766 (1993).
- Dohm, M. R., Hayes, J. P. & Garland, T. Jr *Evolution* **50**, 1688–1701 (1996).
- Gilchrist, G. W. *Evolution* **50**, 1560–1572 (1996).

Competing financial interests: declared none.

brief communications is intended to provide a forum for brief, topical reports of general scientific interest and for technical discussion of recently published material of particular interest to non-specialist readers. Priority will be given to contributions that have fewer than 500 words, 10 references and only one figure. Detailed guidelines are available on *Nature's* website (www.nature.com) or on request from nature@nature.com

Table 1 Correlation between 10 performance measures for 600 world-class decathletes

	Shot putt	Javelin	Pole vault	High jump	Long jump	100 m	110-m hurdles	400 m	1,500 m
Discus	+	+	NS	NS	NS	NS	NS	NS	NS
Shot putt		+	NS	NS	NS	NS	NS	NS	–
Javelin			NS	NS	NS	NS	NS	NS	NS
Pole vault				+	NS	NS	+	NS	NS
High jump					+	NS	+	NS	NS
Long jump						+	+	NS	NS
100 m							+	+	–
110-m hurdles								+	NS
400 m									+

Decathlon competitions consist of: first day, 100 metres, long jump, shot putt, high jump and 400 metres; second day, 110-metre hurdles, discus, pole vault, javelin and 1,500 metres. Partial correlation coefficients were used to adjust correlations between performance pairs for an athlete's overall ability on the basis of the other events (regardless of physiological similarity between traits). For each decathlete, all performance results are taken from their best decathlon score. All data (units in m or in $m\ s^{-1}$) were log-transformed. Significant positive (+) and negative (–) correlations ($P < 0.0001$) were intact after Bonferroni correction (d.f., 590); NS, not significant.

The emerging conceptual framework of evolutionary developmental biology

Wallace Arthur

Ecology Centre, School of Sciences, University of Sunderland, Sunderland SR1 3SD, UK

Over the last twenty years, there has been rapid growth of a new approach to understanding the evolution of organismic form. This evolutionary developmental biology, or 'evo-devo', is focused on the developmental genetic machinery that lies behind embryological phenotypes, which were all that could be studied in the past. Are there any general concepts emerging from this new approach, and if so, how do they impact on the conceptual structure of traditional evolutionary biology? In providing answers to these questions, this review assesses whether evo-devo is merely filling in some missing details, or whether it will cause a large-scale change in our thinking about the evolutionary process.

Evolutionary developmental biology has its origins in the comparative embryology of the nineteenth century, and in particular in the work of von Baer¹ and Haeckel^{2,3}, whose 'laws' were put forward as being generally applicable to the way in which development evolves, regardless of the taxon concerned. Following a quiescent period of almost a century, present-day evo-devo erupted out of the discovery of the homeobox in the early 1980s (refs 4, 5). One principal focus over the last 20 years has been the comparative study of the spatiotemporal expression patterns of developmental genes, both homeobox-containing⁶ (hence coding for transcription factors) and others. Although there are many cases of conservation of expression patterns, there are also cases where these patterns differ markedly between closely related species, even when the gene concerned acts at a very early developmental stage⁷.

We have thus moved from one extreme to the other: from laws that turn out to be, at best, overgeneralizations, to a situation where it almost seems that anything goes, that is, any developmental gene, its expression pattern and the resultant ontogenetic trajectory can evolve in any way. If this were true, no generalizations would be possible, let alone universally applicable laws. However, the search for general patterns is fundamental to science and is not easily suppressed, even when the relevant data set looks very complex, as it currently does in evolutionary developmental biology. There are signs, particularly over the last few years that new general concepts are emerging. I discuss these below, after a brief look at two of evo-devo's foundations.

Comparative embryology and phylogeny

What is the current status of the laws of von Baer and Haeckel? One view is that Haeckel's recapitulation was wrong but von Baer's embryonic divergence essentially right⁸; however, it can be argued⁹ that the pattern that is observed depends on the type of comparison being made. For comparisons among animals of the same level of phenotypic complexity, such as different vertebrate classes, von baerian divergence may be the norm (but see below). However, when comparisons are made between very different levels of complexity, the pattern that emerges is broadly recapitulatory, although only in a very imprecise way, in the sense of recapitulating levels of complexity rather than precise morphological details (Fig. 1). So, taking a broad view, both von Baer and Haeckel captured elements of the truth: evolution leads both to embryonic divergence and, in some lineages, to a lengthening of the ontogenetic trajectory leading to more complex adult phenotypes with greater numbers of cells, their embryos passing through simpler, quasi-ancestral forms. There is, however, an important restriction to this general model.

Von Baer's divergence applies only after the 'phylotypic' stage¹⁰. Earlier development is interspecifically variable, converging to this point of maximum similarity, and only after that diverging again. This 'egg-timer' or 'hourglass' model of development¹¹ renders the situation messier, but should not lead us to abandon the idea of von baerian divergence altogether, especially as the hourglass is a very asymmetric one, with the point of constriction close to the beginning.

Studies of the proposed phylotypic stage in vertebrates reveal more variation than was recognized by either von Baer or Haeckel, partly as a result of screening a wider range of vertebrate species¹². The result is recognition of a phylotypic period, rather than a precise phylotypic stage, and that this period is one of relative, but by no means absolute, conservation. This can be regarded as a specific example of a general phenomenon: the more taxa that are examined, the more any developmental process that initially seemed conserved may be found to vary. This is a strong argument for broad comparative studies.

Such studies rest on another foundation, namely, the explicitly phylogenetic approach to systematics developed by Hennig¹³ between the 1950s and 1970s, plus later refinements of his methodology. This body of work is particularly important, given that the influence of the 'modern synthesis' of the 1930s and 1940s (see below), with its primary focus on evolutionary mechanisms at the population level, resulted in the near-eclipse of phylogenetic considerations from mainstream evolutionary thinking for many years. Most work inspired by the synthesis had no particular antagonism towards phylogeny; its main focus of interest merely lay elsewhere. However, remarks made by a few pro-synthesis biologists reveal their active dismissal of the importance of phylogenetic considerations rather than merely neglect, as, for example, in the Medawars' statement¹⁴ that "nothing of any importance turns on the allocation of one ancestry rather than another".

The importance of the work of Hennig is twofold. First, phylogeny was moved back to a centre-stage position, but without displacing mechanistic approaches from that position, thus setting the scene for a truly combined pattern-and-process approach, even though Hennig focused entirely on pattern. Second, Hennig achieved this by developing the cladistic method for the reconstruction of phylogenetic pattern that is based on a rigorous treatment of data—in terms of shared derived characters—rather than intuitive feelings about systematic relationships arising from accumulated personal experience of particular taxa. Furthermore, his methodology, although initially developed to deal with morphological data, is of very broad applicability and is now in widespread use in the analysis of molecular sequence data, including the sequences of many developmental genes.

Advocates of evolutionary developmental biology are of course interested in the ways in which developmental characters map to phylogenies, as exemplified by studies of comparative echinoderm⁷ and anuran¹⁵ development, and the phylogenetic scope of various characters in the zebrafish model system¹⁶. However, that focus of interest should not be taken to imply that they favour the ontogenetic method of polarizing characters in phylogeny reconstruction¹⁷. In fact, this method has not been widely adopted; rather, the outgroup method has retained favour.

The mapping of developmental processes onto a phylogeny reveals the kinds of alteration in development that we need to explain. Equally, such mapping enables us to exclude from consideration illusory evolutionary changes that never took place, such as the supposed annelid–arthropod transition (see below).

From the homeobox to expression patterns

Evolutionary and developmental biology uncoupled at around 1900, partly due to the latter becoming more experimental and less comparative¹⁸. They stayed largely separate, with some notable exceptions^{19,20}, until about 1980. Their re-integration was prompted by several factors, including the widespread acceptance of phylogenetic systematics (see above), the growth of a more genetic approach to development, especially in *Drosophila*²¹, and the appearance, over a short period around 1980, of a cluster of books on the interface between evolution and development^{8,22,23}. But the single most important factor was the discovery of the homeobox and its widespread phylogenetic conservation^{4,5}.

The importance of homeobox-containing genes (including the *Hox* genes), from an evolutionary perspective, is that they reveal the

existence of a general mechanism underlying the development of morphologically diverse organisms. This constitutes something of a paradox—where does the diversity come from if the genes are highly conserved?—the resolution of which may lie in lower levels of conservation of downstream genes, although this is still largely a presumption. Not only are homeobox genes highly conserved in sequence (especially within the 180-base-pair homeobox region itself), but in many cases their expression patterns are highly conserved also, implying a conservation of function. One of the most enduring visual images of the comparative study of expression patterns of homeobox-containing genes carried out in the 1990s is the picture of ‘stripey’ embryos of various arthropods^{24,25}. These all show the segmentally reiterated transverse bands of expression of the homeobox-containing gene *engrailed*, which reflect its role in determining segment polarity (Fig. 2).

Conservation is, of course, a relative thing. Over long enough periods of evolutionary time, even highly conserved developmental genes evolve, and when they do, important morphological consequences follow. Sometimes altered expression patterns are such as to give more or less of the same thing within a still-conserved overall body plan. An example is divergence in the anteroposterior expression patterns of *Hox* genes in insects and crustaceans, and, associated with this, morphological divergence, for example in the number of leg-bearing segments⁶. Here, the basic design is still the same at a more fundamental level—a segmented arthropod. In other cases, altered expression patterns lead to more fundamental shifts in morphology, and give rise to an entirely distinct body plan. For example, the expression patterns of some homeobox-containing genes in echinoderms²⁶ bear little relation to those in vertebrates,

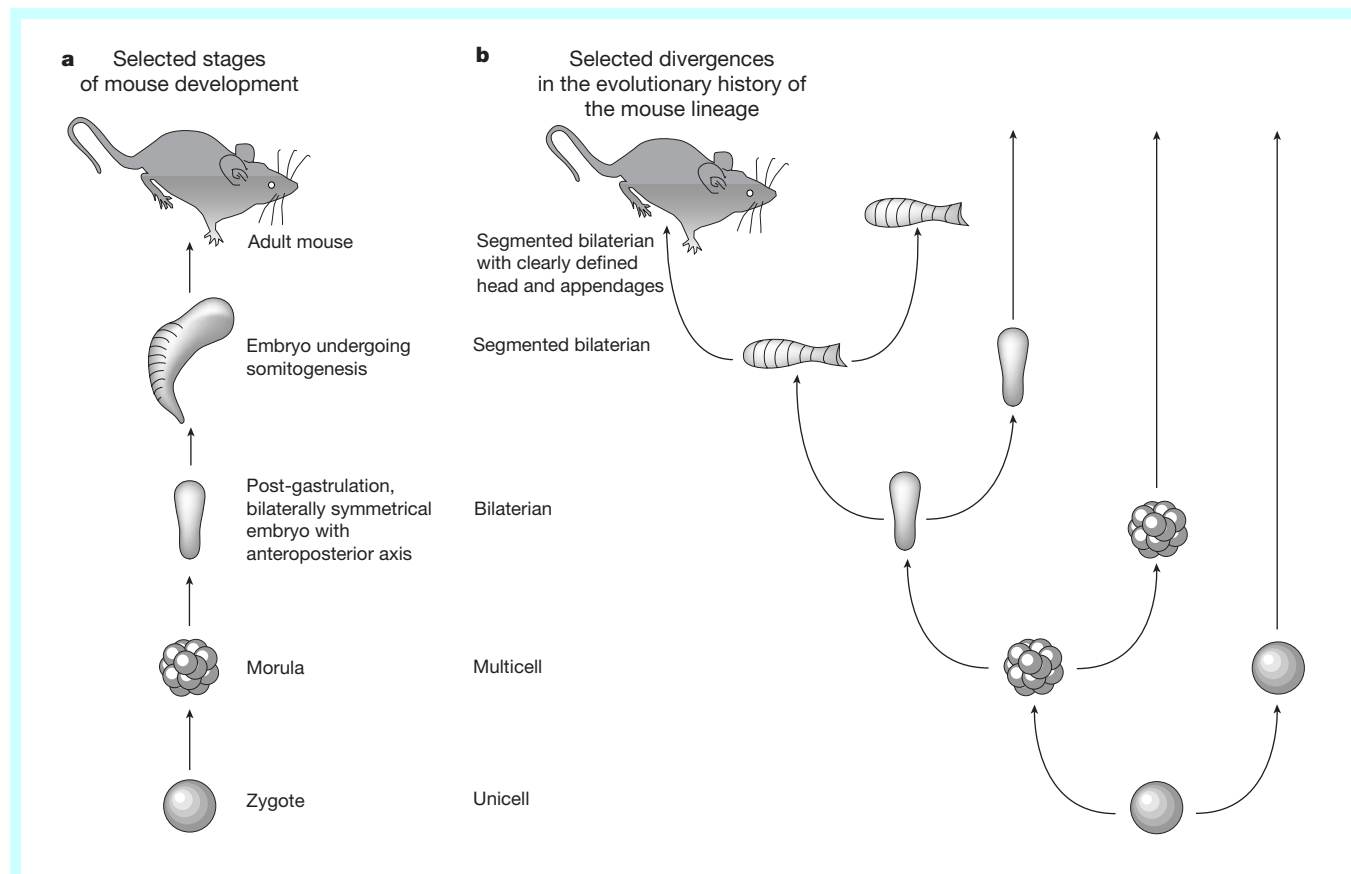


Figure 1 The recapitulatory aspect of the evolution of development in lineages that exhibit increasing phenotypic complexity. The ontogeny of an individual mouse (**a**) recapitulates the broad levels of complexity, but not the precise ancestral forms (either

embryonic or adult), that the mouse lineage passed through in the course of its evolution (**b**).

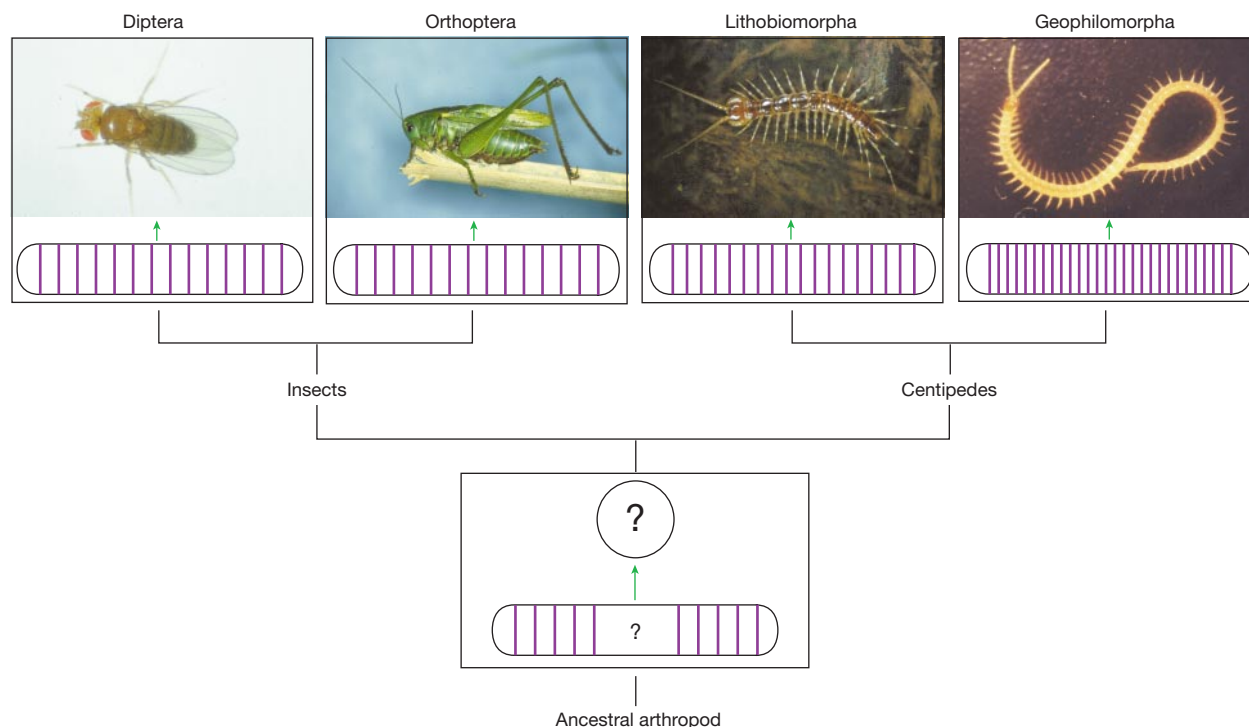


Figure 2 Conservation and change in a developmental mechanism. In all arthropods studied, segments are generated through a cascade of gene interactions, an important component of which is the production of segmental stripes of expression of the segment polarity gene *engrailed*. Starting from an arthropod stem species of unknown embryonic segment number and unknown adult morphology, numerous divergences have occurred in the following: (1) the number of *engrailed* stripes/segments (compare centipedes with each other and with insects); (2) the way in which stripes of expression

are generated in development (roughly simultaneously in flies but in anteroposterior sequence of varying duration in the others); and (3) segment identities, which are controlled separately by *Hox* genes. The source of information for insects is from refs 24 and 25. The centipede *engrailed* images were based, at the time of writing, on extrapolation from the known segmental expression patterns of this gene in other arthropods. Recent work has confirmed that such extrapolation is justified (C. Hughes, T. Kaufman and C. Kettle, unpublished data).

as perhaps might be expected given the fundamental gap between bilateral and pentaradial symmetry. The dorsoventral axis inversion that accompanied the divergence of protostomes and deuterostomes^{27,28} is intermediate between the above two examples.

Evo-devo and general evolutionary theory

How do all these studies impact on evolutionary biology, a conceptually driven discipline? Perhaps comparative studies of expression patterns are just 'filling in the details'. After all, if morphology evolves, these things that are the underlying causes of morphology must also be evolving. So it might be argued that the accumulating data of evolutionary developmental biology are of no particular consequence for the overall structure of evolutionary theory. However, I will now put forward a rather different view.

As is widely acknowledged, development had a minor role both in the work of Darwin²⁹ and in the establishment of the modern synthesis of the mid-twentieth century^{30–32}. One reason is that even as late as the completion of the synthesis, developmentalists were conceptually isolated from geneticists and evolutionists, for whom the gene had by then assumed a central importance. Developmental concepts of the day, such as induction, gradients and fields, were hard to visualize from a genetic perspective, and thus hard to accommodate within a gene-centred view. Today all that has changed, and developmental genetics is a very large part of developmental biology. Now that the barrier has disappeared, does the new developmental perspective have a conceptual contribution to make to our overall understanding of evolution?

In fact, several groups of concepts have emerged recently from evolutionary developmental biology. Some are new, whereas others

are resurrections or elaborations of much older concepts (going back to von Baer and beyond). These concepts are listed in Table 1, along with key references. There are too many to discuss in detail so it will be necessary to be selective. Below I discuss the first two groups of concepts in Table 1: one centred around the idea of developmental reprogramming, the other around the idea of gene co-option.

Developmental reprogramming and bias

Given that development has long been the 'missing link' of evolutionary theory, where and how does that link fit, in terms of conceptual structure? There is a very simple answer to this question (Table 2). Evolution is often portrayed as an interplay between mutation and selection, with the former providing a supply of variation, and the latter acting as a fitness-based sieve. But this picture has a major flaw, because mutation provides new genes, whereas selection acts not on genes but on phenotypes. The missing link, then, is how we get from altered gene to new phenotype. This is the process of developmental reprogramming³³—that is, a mutationally driven change in something that is itself a state of change. If a particular ontogeny is represented by a trajectory through multi-dimensional phenotypic space, then after reprogramming we have a different trajectory. Many types and degrees of difference are possible.

The term reprogramming should be interpreted with care. Looked at in one way, development is programmed by genes. But this is too limited a view³⁴. There is a complementary process, the epigenetic programme, through which genes are controlled by developmental agents of diverse kinds, including transcription

Table 1 Key concepts of evolutionary developmental biology

General theme	Specific concept	Selected references
The nature of developmental variation	Developmental reprogramming	33, 72
	Developmental bias/mutation bias	41, 42
	Developmental constraint	43, 44
	Developmental drive	45
	Developmental reaction norms	36, 37, 71
Re-use of developmental genes in evolution	Co-option	55, 56, 58, 63
	Exaptation	54, 55
	Cassettes of developmental genes	56, 57
	Paramorphism	57
Aspects of evolutionary conservation	Body plans	9
	Evolutionarily stable configurations	73
	Phyloclonal stage/period	10–12
	Zootype	74
	Homology	75–77
Factors promoting evolutionary change	Modularity	78
	Dissociability	78
	Evolvability	79
	Duplication and divergence (of body parts and of genes)	78, 80
Evo-devo aspects of natural selection	Developmental/internal selection	81
	Generative entrenchment/burden	82, 83
	Co-adaptation of developmental processes	9
	Unmasking of hidden variation	84
	Genetic assimilation	20

The concepts listed, both old and recent, are grouped according to the general theme to which they relate. Selected references are indicated. The list is not intended to be exhaustive.

factors and secreted morphogens. The delayed entry of the zygotic genome into the control of early development in most animals³⁵ reveals that the very beginnings of the developmental process are controlled epigenetically. Finally, there is an ecological aspect to the overall programme, because the ontogenetic trajectory is often influenced by environmental factors (the reaction norm³⁶), sometimes in an abrupt way, as in the case of aquatic plants that make different kinds of leaves above and below water (heterophylly³⁷), or the insects that produce winged and wingless forms at different population densities. More often, the whole ontogenetic trajectory is deflected in a continuously variable way in response to environmental conditions.

What constitutes developmental reprogramming thus becomes clear. It is a mutation-based, and thus inherited change in the overall genetic/epigenetic/ecological programme through which we get from genome to phenome. Importantly, it is a theory-neutral process, to which even the staunchest neo-darwinian is unlikely to object. Reprogramming becomes controversial only if we propose that it exhibits intrinsic biases (see below), which may provide an additional evolutionary mechanism. Reprogramming can be examined at many different levels, from the alteration of a developmental gene's product through its ontogenetic consequences to (in some cases) its ultimate effects on the adult phenotype. At any one level, there are four possibilities: changes in timing, spatial distribution, quantity and type; respectively heterochrony^{38–40}, heterotopy^{39,40}, heterometry³³ and heterotypy³³. However, a change of one of these kinds at the molecular level may well give rise to other kinds of change at the phenotypic level.

The conclusion is that intraspecific evolutionary change in morphology requires, at minimum, the three processes of mutation,

reprogramming and selection, whereas trans-specific evolution requires, in addition, reproductive isolation. The only barrier to the general inclusion of reprogramming as a logically necessary component of the evolutionary process is a passive one, namely inertia. There is no element of antagonism in the sense that there is no branch of evolutionary thought that is anti-development *per se*.

The picture changes radically if it is proposed that reprogramming is, at least in some cases, systematically biased, in that mutation more readily produces changes in certain directions than others, including the extreme case of some directions being apparently 'prohibited'. Such a state of affairs in general has been referred to as mutation bias⁴¹ or developmental bias ('evolution biased by development'⁴²). Negative biases, both relative and absolute, constitute constraint^{43,44}, whereas positive biases have recently been termed developmental drive⁴⁵ (quite distinct from meiotic drive⁴⁶, molecular drive⁴⁷ and dominance drive⁴⁸). Proposals that these biases can potentially lead to the direction of evolutionary change being determined by developmental dynamics as well as by population dynamics are in contrast with the historical thrust of darwinism and neo-darwinism, that the direction of change is determined exclusively by selection^{49,50}.

The main problem facing proponents of a directional evolutionary role for developmental bias is that evidence for this is both very limited at present and harder to acquire than evidence for selection. The predominance of certain leaf-arrangement⁵¹ and floral-symmetry⁵² patterns in angiosperms have been proposed as examples of bias-led evolution. The fact that all 3,000 or so species of centipede have odd numbers of leg-bearing segments (from 15 to 191) also suggests developmental bias⁵³, and in this case an alternative selective explanation is highly implausible.

Table 2 Processes causing evolutionary change at four levels of biological organization

Level	Initial state	Process causing change	Altered state
Gene	Gene making morphogenetic protein	Mutation	New version of gene and new protein
Organism	Organism with particular ontogenetic trajectory	Developmental reprogramming	New ontogenetic trajectory
Population	Population with certain relative frequencies of 'old' and 'new' ontogenies	Natural selection	Population with altered frequencies of the two ontogenies
Species	Genetically distinct populations	Reproductive isolation	Pair or group of daughter species

Attention is focused here on the main process at each level; this should not be taken as a denial of a role for additional processes (for example, genetic drift at the population level). The source of information for this table is ref. 33.

Are such examples exceptions to a general rule that selection on its own determines evolutionary directionality, or are they an indication of a general but as yet largely undiscovered role for developmental bias? This is an entirely open and very important question, which is likely to be a major focus of future studies. These should include both further studies of unusual phylogenetic distributions of character states, including excesses, deficiencies and absences of particular states for which selective explanations seem implausible, and studies that attempt to quantify the tendency of mutations in model-system organisms to produce reprogramming in some directions rather than others. Considerable care will be needed in the design of these studies, however, if explanations other than developmental bias are to be excluded.

Co-option, exaptation and paramorphism

A central concept that is emerging from comparative studies of developmental genes and their interactions is the co-option of these for different functional roles as evolution proceeds. This is essentially a molecular version of the concept of exaptation^{54,55}, in which a structure evolved for one adaptive reason is later 'exapted' (or co-opted) for some other role. Variations on this theme are: the idea that genes are co-opted not individually but as interacting cassettes⁵⁶, and the idea that co-option accompanies and helps to cause the formation of new structures (paramorphism⁵⁷; see below) rather than occurring afterwards to refine them.

The 'pattern-before-process' argument that emerged from cladistics (see above) is particularly important in considering the possible co-option of developmental genes⁵⁸. For example, overt metameric segmentation is pervasive in three high-level taxonomic groups: annelids, arthropods and chordates. It was originally thought that annelids and arthropods were sister groups⁵⁹, and thus that segmentation had originated twice—in an annelid/arthropod common ancestor and in a primitive chordate. More recently, it was suggested that the bilaterian stem group (Urbilateria) was segmented⁶⁰, and that segmentation had thus originated only once, although this seems implausible given the large number of secondary losses it implies in unsegmented phyla. But with the recognition of the superphyletic groups Ecdysozoa and Lophotrochozoa⁶¹ in addition to Deuterostomia, it is now generally thought that segmentation arose on three separate occasions. If this is true, why is some of the developmental-genetic machinery underlying segmentation shared between two or all of the three metamerous phyla? For example, why is there segmentally reiterated expression of an *engrailed* homologue in some chordates⁶², resembling—at least in broad terms—the pattern that is well known in arthropods^{24,25}? Is this a case of independent co-option?

Before attempting to answer this question, it is worth noting the parallel situation in limb formation⁶³. Arthropod and vertebrate limbs are not regarded as homologous, yet the formation of both is characterized by expression of *Distal-less* (*Dll*) in the prospective limb-tip region. And indeed, clearly non-homologous outgrowths in other animals, for example the tube feet of echinoderms, are also characterized by *Dll* expression. So again, does this represent independent co-option of homologous developmental genes for similar topographic roles in different evolutionary lineages? Is such co-option not rather improbable?

A possibility that is gaining favour^{58,63} is that the genes concerned may have had a function in a common ancestor that was of the same general kind as now observed, but in a different developmental context. For example, in an ancestral bilaterian that had neither segments nor legs, there may nevertheless have been subdivision of some internal body part (*engrailed*) or some rudimentary outgrowth (*Distal-less*). If, following some subsequent lineage divergence, two or more descendant species independently become segmented, or independently develop limbs, the easiest way to do so may be to make use of a developmental system already in existence.

This is unlikely to happen on a piecemeal basis, with one gene at a time being co-opted from the original to the new function. Why keep using *engrailed* in segmentation and *Distal-less* in limb outgrowth? It is not just the fact that the products of these genes and of their upstream controllers have gene-regulatory roles that is important; rather it is the specificity of those roles. If a gene whose product acts at a particular point in an interaction cascade can be expressed at an ectopic location, then the whole cassette downstream of that may be expressed in the new location too. Notably, *engrailed* and *Distal-less* are both fairly far downstream in their respective pathways, so their activation in new locations may be just such a downstream consequence of the ectopic activation of their upstream controllers. However, this apparently neat model may be deceptive (see Box 1).

This line of argument has been taken one stage further in relation to limbs. It has been argued that the initial formation of proto-limbs in an originally limbless lineage is not something that will occur before, and in isolation from, the co-option of a cassette of outgrowth-inducing genes. Rather, these things will happen together⁵⁷, with the result that limbs can be thought of as 'axis paramorphs', that is, scaled-down, rotated, and simplified (for example, lacking endoderm) versions of the main anteroposterior body axis.

The key question now becomes: what kind of mutation causes the initially ectopic expression of cassettes of developmental genes to occur in a spatiotemporal pattern that affords some possibility of functional improvement and thus of being favoured by selection? Although we cannot yet answer this question, it is interesting to consider a possible link with developmental bias. The best way to do this is to contrast the evolutionary initiation of limb formation with an alternative case of developmental evolution often dealt with by quantitative geneticists and evolutionary ecologists, namely evolution of body size. There are two important differences between these two systems. Increased or decreased body size is a rather generalized change, and there is almost always heritable variation in a population around the mean size. Thus the situation is perhaps not one in which developmental bias is likely to have a role in determining the direction of evolutionary change; rather, this is set by selection. But the ectopic activation of whole cassettes of developmental genes involved in outgrowth formation in a particular spatiotemporal pattern is both much more specific and much less likely to be based on routinely occurring intrapopulation variation. Both of these features make it much more probable that developmental bias will have a role here. Indeed, we can extrapolate from this comparison and argue that developmental bias is most likely to have an important role when major evolutionary innovations are taking place.

Intraspecific origins

Although much of evolutionary developmental biology concerns large-scale comparisons, often between classes and phyla, it seems probable that all evolutionary changes ultimately begin as intraspecific variation—whether routine, as in the case of body size, or exceptional, as in the case of the origin of limbs. This variation occurs within particular populations living in particular environments. Thus, to investigate the evolutionary origins of interspecific differences, either in developmental genes or in the corresponding ontogenies and adult phenotypes, it is desirable to examine intraspecific variation in the same genes, ontogenies or phenotypes. So far, work in this area is rather limited. Nevertheless, two different approaches can already be distinguished. The first is to take a classic evo-devo model system and to add a population dimension. The second is to take a classic ecological genetics model system and to add a developmental dimension. We can now look at examples of these in turn.

Segmentation provides a good system for the evo-devo to population approach. An important finding here is that one of the *Hox* genes determining segment identity, *Ultrabithorax*, exhibits intraspecific variation, in the form of polymorphism, within

Box 1

Developmental pathways and their evolution

The concept of an interaction pathway — and the group of genes that encode its components — as comprising a sort of developmental cassette⁶⁶ that can be treated as a unit of evolutionary change is an attractive one. Different kinds of evolutionary process involving such units can be considered, including their divergence in separate lineages after speciation and their co-option for a new developmental purpose within a lineage, possibly coincident with the appearance of new morphological structures such as limbs (paramorphism⁵⁷; see text). However, development is a complex process, and evolution, because of its stochastic and historical aspects, works in a messy way. Therefore, we need to be cautious in attempting to apply neat concepts like co-option of whole cassettes. Here, I outline some of the relevant complexities.

Complexities relating to development itself. A single developmental gene can have a very complex pattern of expression⁸⁵. This may be aided by different promoters or other controlling elements regulating expression in different tissues or at different times, such as occurs in the *Pax6* gene, which is involved in eye development⁸⁶. When we consider a whole pathway, another level of complexity appears. In the *Drosophila* segmentation pathway, the four groups of genes involved — maternal, gap, pair rule and segment polarity — were well characterized more than a decade ago⁸⁷. It now seems that there are about ten genes per group, and so about forty in total. But the pattern of interaction is far from being a linear cascade. Some of the genes fall into two groups. For example, *hunchback*, which is normally thought of as a gap gene, is expressed both maternally and zygotically. And a newly discovered gene, *lilliputian*, seems to be maternal, yet mutants exhibit a pair-rule phenotype⁸⁸. Also, even without this complication, the pair-rule group of genes is a double group in the sense that some of its members are 'primary' and some 'secondary'. Furthermore, as well as interactions between groups, there are interactions within them, and these take a variety of forms. At a smaller scale, all interactions are likely to involve variation in such parameters as the stability of RNA and protein molecules, although computer modelling⁸⁹ suggests that the overall result, in terms of the pattern of expression of segment polarity genes like *engrailed*, may be robust in the sense of being broadly unaffected by such variation.

No such pathway exists in isolation from the rest of the developing organism. This makes the use of the terms upstream and downstream problematic as descriptors of developmental genes and their products. It is essential to recognize that these are relative terms. For example, *engrailed* is often seen as a downstream gene because it falls into the segment polarity group. But a vast amount of morphological detail lies between the rudimentary proto-segments that exist when

engrailed stripes first appear, and their final, fully elaborated form. So, from another perspective, *engrailed* is an upstream gene. The same argument can be used regarding *Distal-less*, which comes at the 'end' of a series of interactions (with *decapentaplegic* and *wingless* being immediately upstream⁹⁰), but whose product goes on to control downstream processes. A converse argument can be made for *Pax6*, which is normally considered to be upstream: it is downstream from another perspective⁹¹.

Complexities that emerge from comparative studies.

Broad-scale comparative studies extending beyond model systems to encompass many taxa usually start with, or have a bias towards, a particular gene in the pathway concerned. Often, this is for methodological reasons, such as the use of the monoclonal antibody 4D9 in detecting homologues of *engrailed*^{24,25}. Given the widespread conservation of the segmental expression of this gene, it would have seemed reasonable to predict an equal or even greater degree of conservation of those genes upstream of it. This would be the pattern expected on the basis of a genetic version of von Baer's laws. So far, however, this prediction has not been borne out. One of the earliest-acting genes in the pathway — the anterior determinant *bicoid* — seems to be a relatively recent evolutionary innovation found only in a smallish clade⁹². Furthermore, both *bicoid* itself and its interaction with *hunchback* seem to be capable of quite rapid evolution^{93,94}. It is also possible that the pair-rule component of the pathway is not highly conserved. However, the details of this pathway are still being investigated. It was initially thought that in the orthopteran *Schistocerca* (an insect with short-germ development) there was no two-segment periodicity of expression of homologues of the *Drosophila* pair-rule genes⁹⁵; however, such expression has recently been demonstrated⁹⁶. A similar expression pattern of pair-rule homologues has been confirmed for the beetle *Tribolium*⁹⁷, which is phylogenetically intermediate between *Schistocerca* and *Drosophila*.

Caution, but not despair. An understanding of the extraordinary complexity of ecosystems did not deter Darwin from his belief in a single, simple idea (natural selection) that we still regard as the central, 'external' principle of evolution, that is, the main mechanism responsible for adaptation of organisms to their environment. Our growing understanding of the many complexities of development similarly should not deter us from searching for simple general concepts that relate to the 'internal' aspect of evolution, including the ways in which developmental pathways evolve (such as an evolutionary trend towards hierarchical patterns of interaction^{98,99}), and the ways in which the internal co-adaptation of organisms is maintained and enhanced.

outbred populations of *Drosophila melanogaster*⁶⁴. This connects with a possible role for *Ultrabithorax* in the evolution of morphological differences between congeneric species⁶⁵ (*D. melanogaster* and *D. simulans*). Also, in relation to genes determining segment number rather than identity, a few groups of arthropods exhibit intraspecific variation in segment number, based on different numbers of *engrailed* expression stripes, in turn generated by variation in upstream genes. In one case (the centipede *Strigamia maritima*) comparative studies of populations living at different latitudes reveal a cline⁶⁶, and consequently suggest the action of selection.

Naturally occurring variation in a plant developmental gene provides another example of the evo-devo to population approach. Populations of *Arabidopsis thaliana* show considerable molecular variation at the CAULIFLOWER locus, which encodes a transcription factor (of the MADS-box family⁶⁷) that controls the development of inflorescence meristems⁶⁸. The pattern of variation found suggests the action of natural selection. Also, it appears that the artificial selection used by farmers to produce the domestic varieties

of *Brassica oleracea* (cauliflower and broccoli) caused changes at the CAULIFLOWER locus⁶⁹, and presumably at other loci as well.

Butterfly wing pigmentation patterns provide a good system for the ecological genetics to evo-devo approach. Early studies (for example, on mimicry⁷⁰) had both ecological and transmission genetics dimensions, but lacked information on developmental mechanisms. Recent studies of eyespots in the butterfly *Bicyclus anynana*⁷¹ have added this missing dimension. One notable result to emerge from these studies is that the gene *Distal-less*, a familiar component of limb formation, is also involved in the generation of eyespots. Selection experiments reveal that this system is readily modifiable, and it looks as though the cassette concept may apply here, with changes in the expression patterns of interacting genes.

Towards a synthesis?

Thus far, I have discussed two of the five groups of concepts in Table 1 in detail, but will now briefly address the connections between all five of these groups, in addition to looking at approaches that might be most productive for the future.

The third, fourth and fifth groups of concepts in Table 1 relate to the interaction between conservation and change. Unsurprisingly, certain features favour one, while their opposites favour the other. For example, an early stage of development in which the embryo is developing as a coordinated whole (for example, the vertebrate neurula) is highly entrenched, resistant to selection, and thus widely conserved, although there are many exceptions to this pattern, especially in multi-stage life histories with free-living larval forms. In contrast, later stages (for example, limb buds) are often quasi-autonomous modules of enhanced evolvability, although caution is necessary because some of the genes involved may also have other developmental roles. When selectively driven change is possible, it is sometimes aided by mechanisms for the unmasking of variation; the selection itself may relate to external adaptation, internal co-adaptation, or a mixture of the two. In all cases of evolutionary change in ontogeny, developmental reprogramming is a necessary part of the process, and one way of achieving it is co-option of existing developmental genes for new roles. Developmental bias may interact with selection to govern the direction of change, although this is just an intriguing, but largely untested possibility.

With regard to future work, three approaches stand out as being particularly desirable. First we need to develop rigorous tests for developmental bias, as noted above. Second, we need some thorough case studies of downstream developmental genes to contrast with their upstream counterparts. (Even the most downstream of those that I have used as examples are quite far upstream in relation to the totality of development; see Box 1.) We need to ask whether both DNA sequences and gene functions are less conserved in downstream genes, especially those that can be shown to have no cryptic early role in addition to their late one. Third, we need to focus experimental work on the important concepts of co-option, cassettes and paramorphism. This can include both high-level comparative studies and studies focusing on mutations (for example, in *Drosophila*) that show ectopic expression of developmental genes that could provide a basis for the establishment of new roles. In other words, we should attack the problem at both ends—its origins in terms of mutation and reprogramming within species, and its long-term results, manifested as accumulated evolutionary divergence over hundreds of millions of years. □

1. Von Baer, K. E. *Über Entwicklungsgeschichte der Tiere: Beobachtung und Reflexion* (Borntäger, Königsberg, 1828).
2. Haeckel, E. *Generelle Morphologie der Organismen* (Georg Reimer, Berlin, 1866).
3. Haeckel, E. *The Evolution of Man: a Popular Exposition of the Principal Points of Human Ontogeny and Phylogeny* (Appleton, New York, 1896).
4. Scott, M. P. & Weiner, A. J. Structural relationships among genes that control development: sequence homology between the *Antennapedia*, *Ultrathorax* and *fushi tarazu* loci of *Drosophila*. *Proc. Natl Acad. Sci. USA* **81**, 4115–4119 (1984).
5. McGinnis, W., Garber, R. L., Wirz, J., Kuroiwa, A. & Gehring, W. J. A homologous protein-coding sequence in *Drosophila* homeotic genes and its conservation in other metazoans. *Cell* **37**, 403–408 (1984).
6. Averof, M. & Akam, M. *Hox* genes and the diversification of insect and crustacean body plans. *Nature* **376**, 420–423 (1995).
7. Wray, G. A. & Bely, A. E. In *The Evolution of Developmental Mechanisms* (eds Akam, M., Holland, P., Ingham, P. & Wray, G.). *Development* (Suppl.) (Company of Biologists, Cambridge, 1994).
8. Gould, S. J. *Ontogeny and Phylogeny* (Harvard Univ. Press, Cambridge, Massachusetts, 1977).
9. Arthur, W. *The Origin of Animal Body Plans: a Study in Evolutionary Developmental Biology* (Cambridge Univ. Press, Cambridge, 1997).
10. Sander, K. In *Development and Evolution* (eds Goodwin, B. C., Holder, N. & Wylie, C. C.) 137–159 (Cambridge Univ. Press, Cambridge, 1983).
11. Duboule, D. In *The Evolution of Developmental Mechanisms* (eds Akam, M., Holland, P., Ingham, P. & Wray, G.). *Development* (Suppl.) (Company of Biologists, Cambridge, 1994).
12. Richardson, M. K. *et al.* There is no highly conserved embryonic stage in the vertebrates: implications for current theories of evolution and development. *Anat. Embryol.* **196**, 91–106 (1997).
13. Hennig, W. *Phylogenetic Systematics* (Univ. Illinois Press, Urbana, 1966).
14. Medawar, P. B. & Medawar, J. S. *The Life Science: Current Ideas of Biology* (Wildwood House, London, 1977).
15. Maglia, A. M., Pugener, L. A. & Trueb, L. Comparative development of anurans: using phylogeny to understand ontogeny. *Am. Zool.* **41**, 538–551 (2001).
16. Metscher, B. D. & Ahlberg, P. E. Zebrafish in context: uses of a laboratory model in comparative studies. *Dev. Biol.* **210**, 1–14 (1999).
17. Nelson, G. J. Ontogeny, phylogeny, paleontology and the biogenetic law. *Syst. Zool.* **27**, 324–345 (1978).
18. Roux, W. The problems, methods and scope of developmental mechanics. *Biol. Lect. Mar. Biol. Lab., Woods Hole* 149–190 (Ginn, Boston, 1894).
19. De Beer, G. R. *Embryos and Ancestors* (Clarendon, Oxford, 1940).
20. Waddington, C. H. *The Strategy of the Genes* (Allen & Unwin, London, 1957).
21. Nüsslein-Volhard, C. & Wieschaus, E. Mutations affecting segment number and polarity in *Drosophila*. *Nature* **287**, 795–801 (1980).
22. Raff, R. A. & Kaufman, T. C. *Embryos, Genes and Evolution: the Developmental Genetic Basis of Evolutionary Change* (Macmillan, New York, 1983).
23. Arthur, W. *Mechanisms of Morphological Evolution: a Combined Genetic, Developmental and Ecological Approach* (Wiley, Chichester, 1984).
24. Patel, N. H. Developmental evolution: insights from studies of insect segmentation. *Science* **266**, 581–590 (1994).
25. Patel, N. H. In *The Evolution of Developmental Mechanisms* (eds Akam, M., Holland, P., Ingham, P. & Wray, G.). *Development* (Suppl.) (Company of Biologists, Cambridge, 1994).
26. Lowe, C. J. & Wray, G. A. Radical alterations in the roles of homeobox genes during echinoderm evolution. *Nature* **389**, 718–721 (1997).
27. Geoffroy Saint-Hilaire, E. Considérations générales sur la vertèbre. *Mem. Mus. Hist. Nat.* **9**, 89–119 (1822).
28. Holley, S. A. *et al.* A conserved system for dorsal-ventral patterning in insects and vertebrates involving *sog* and *chordin*. *Nature* **376**, 249–253 (1995).
29. Darwin, C. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* (John Murray, London, 1859).
30. Dobzhansky, T. *Genetics and the Origin of Species* (Columbia Univ. Press, New York, 1937).
31. Mayr, E. *Systematics and the Origin of Species* (Columbia Univ. Press, New York, 1942).
32. Simpson, G. G. *Tempo and Mode in Evolution* (Columbia Univ. Press, New York, 1944).
33. Arthur, W. The concept of developmental reprogramming and the quest for an inclusive theory of evolutionary mechanisms. *Evol. Dev.* **2**, 49–57 (2000).
34. Robert, J. S. Interpreting the homeobox: metaphors of gene action and activation in development and evolution. *Evol. Dev.* **3**, 287–295 (2001).
35. Davidson, E. H. *Gene Activity in Early Development* 3rd edn (Academic, Orlando, 1986).
36. Schlichting, C. D. & Pigliucci, M. *Phenotypic Evolution: a Reaction Norm Perspective* (Sinauer, Sunderland, 1998).
37. Wells, C. L. & Pigliucci, M. Adaptive phenotypic plasticity: the case of heterophyly in aquatic plants. *Persp. Plant Ecol. Evol. Syst.* **3**, 1–18 (2000).
38. McKinney, M. L. & McNamara, K. J. *Heterochrony: the Evolution of Ontogeny* (Plenum, New York, 1991).
39. Hall, B. K. *Evolutionary Developmental Biology* 2nd edn (Kluwer, Dordrecht, 1999).
40. Zelditch, M. L. & Fink, W. L. Heterochrony and heterotopy: stability and innovation in the evolution of form. *Paleobiology* **22**, 241–254 (1996).
41. Yampolsky, L. Y. & Stoltzfus, A. Mutation bias as an orienting factor in selective evolution. *Evol. Dev.* **3**, 73–83 (2001).
42. Raff, R. A. Evo-devo: the evolution of a new discipline. *Nature Rev. Genet.* **1**, 74–79 (2000).
43. Maynard Smith, J. *et al.* Developmental constraints and evolution. *Q. Rev. Biol.* **60**, 265–287 (1985).
44. Gould, S. J. & Lewontin, R. C. The Spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proc. R. Soc. Lond. B* **205**, 581–598 (1979).
45. Arthur, W. Developmental drive: an important determinant of the direction of phenotypic evolution. *Evol. Dev.* **3**, 271–278 (2001).
46. Geliva, E. A. Meiotic drive in the sex chromosome system of the varying lemming, *Dicrostomys torquatus* Pall. (Rodentia: Microtinae). *Heredity* **59**, 383–389 (1987).
47. Dover, G. A. Molecular drive: a cohesive mode of species evolution. *Nature* **299**, 111–117 (1982).
48. Mallet, J. Hybrid zones of *Heliconius* butterflies in Panama and the stability and movement of warning colour clines. *Heredity* **56**, 191–202 (1986).
49. Fisher, R. A. *The Genetical Theory of Natural Selection* (Clarendon, Oxford, 1930).
50. Ford, E. B. *Ecological Genetics* 3rd edn (Chapman & Hall, London, 1971).
51. Goodwin, B. *How the Leopard Changed its Spots: The Evolution of Complexity* (Weidenfeld & Nicolson, London, 1994).
52. Donoghue, M. J. & Ree, R. H. Homoplasy and developmental constraint: a model and an example from plants. *Am. Zool.* **40**, 759–769 (2000).
53. Arthur, W. & Farrow, M. The pattern of variation in centipede segment number as an example of developmental constraint in evolution. *J. Theor. Biol.* **200**, 183–191 (1999).
54. Gould, S. J. & Vrba, E. S. Exaptation—a missing term in the science of form. *Paleobiology* **8**, 4–15 (1982).
55. Chipman, A. D. Developmental exaptation and evolutionary change. *Evol. Dev.* **3**, 299–301 (2001).
56. Tabin, C. J., Carroll, S. B. & Panganiban, G. Out on a limb: parallels in vertebrate and invertebrate limb patterning and the origin of appendages. *Am. Zool.* **39**, 650–663 (1999).
57. Minelli, A. Limbs and tail as evolutionarily diverging duplicates of the main body axis. *Evol. Dev.* **2**, 157–165 (2000).
58. Abouheif, E. *et al.* Homology and developmental genes. *Trends Genet.* **13**, 432–433 (1997).
59. Clark, R. B. *Dynamics in Metazoan Evolution: the Origin of the Coelom and Segments* (Clarendon, Oxford, 1964).
60. De Robertis, E. M. The ancestry of segmentation. *Nature* **387**, 25–26 (1997).
61. Aguinaldo, A. M. A. *et al.* Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* **387**, 489–493 (1997).
62. Holland, L. Z., Kene, M., Williams, N. A. & Holland, N. D. Sequence and embryonic expression of the amphioxus *engrailed* gene (*AmphiEn*): the metameric pattern of transcription resembles that of its segment-polarity homolog in *Drosophila*. *Development* **124**, 1723–1732 (1997).
63. Panganiban, G. *et al.* The origin and evolution of animal appendages. *Proc. Natl Acad. Sci. USA* **94**, 5162–5166 (1997).
64. Gibson, G. & Hogness, D. S. Effect of polymorphism in the *Drosophila* regulatory gene *Ultrathorax* on homeotic stability. *Science* **271**, 200–203 (1996).
65. Stern, D. L. A role of *Ultrathorax* in morphological differences between *Drosophila* species. *Nature* **396**, 463–466 (1998).
66. Kettle, C. & Arthur, W. Latitudinal cline in segment number in an arthropod species, *Strigamia maritima*. *Proc. R. Soc. Lond. B* **267**, 1393–1397 (2000).
67. Ng, M. & Yanofsky, M. F. Function and evolution of the plant MADS-box gene family. *Nature Rev. Genet.* **2**, 186–193 (2001).
68. Purugganan, M. D. & Suddith, J. I. Molecular population genetics of the *Arabidopsis* CAULIFLOWER

- regulatory gene: Nonneutral evolution and naturally occurring variation in floral homeotic function. *Proc. Natl Acad. Sci. USA* **95**, 8130–8134 (1998).
69. Purugganan, M. D., Boyles, A. L. & Suddith, J. I. Variation and selection at the *CAULIFLOWER* floral homeotic gene accompanying the evolution of domesticated *Brassica oleracea*. *Genetics* **155**, 855–862 (2000).
 70. Turner, J. R. G. Butterfly mimicry: the genetical evolution of an adaptation. *Evol. Biol.* **10**, 163–226 (1977).
 71. Brakefield, P. M. *et al.* Development, plasticity and evolution of butterfly eyespot patterns. *Nature* **384**, 236–242 (1996).
 72. Roth, G. & Wake, D. B. Trends in the functional morphology and sensorimotor control of feeding behaviour in salamanders: an example of the role of internal dynamics in evolution. *Acta Biotheor.* **34**, 175–192 (1985).
 73. Wagner, G. P. & Schwenk, K. Evolutionarily stable configurations: functional integration and the evolution of phenotypic stability. *Evol. Biol.* **31**, 155–217 (2000).
 74. Slack, J. M. W., Holland, P. W. H. & Graham, C. F. The zootype and the phylotypic stage. *Nature* **361**, 490–492 (1993).
 75. Owen, R. *On the Archetype and Homologies of the Vertebrate Skeleton* (John van Voorst, London, 1848).
 76. Hall, B. K. (ed.) *Homology: the Hierarchical Basis of Comparative Biology* (Academic, San Diego, 1994).
 77. Mindell, D. P. & Meyer, A. Homology evolving. *Trends Ecol. Evol.* **16**, 434–440 (2001).
 78. Raff, R. A. *The Shape of Life: Genes, Development and the Evolution of Animal Form* (Chicago Univ. Press, Chicago, 1996).
 79. Kirschner, M. & Gerhart, J. Evolvability. *Proc. Natl Acad. Sci. USA* **95**, 8420–8427 (1998).
 80. Holland, P. W. H. Gene duplication: past, present and future. *Sem. Cell Dev. Biol.* **10**, 541–547 (1999).
 81. Whyte, L. L. *Internal Factors in Evolution* (Tavistock, London, 1965).
 82. Wimsatt, W. C. in *Integrating Scientific Disciplines* (ed. Bechtel, W.) (Martinus-Nijhoff, Dordrecht, 1986).
 83. Riedl, R. *Order in Living Organisms: a Systems Analysis of Evolution* (Wiley, Chichester, 1978).
 84. Rutherford, S. L. & Lindquist, S. Hsp90 as a capacitor for morphological evolution. *Nature* **396**, 336–342 (1998).
 85. Salser, S. J. & Kenyon, C. A. C. *elegans Hox* gene switches on, off, on and off again to regulate proliferation, differentiation and morphogenesis. *Development* **122**, 1651–1661 (1996).
 86. Xu, P.-X. *et al.* Regulation of *Pax6* expression is conserved between mice and flies. *Development* **126**, 383–395 (1999).
 87. Ingham, P. W. The molecular genetics of embryonic pattern formation in *Drosophila*. *Nature* **335**, 25–34 (1988).
 88. Tang, A. H., Neufeld, T. P., Rubin, G. M. & Mueller, H.-A. J. Transcriptional regulation of cytoskeletal functions and segmentation by a novel maternal pair-rule gene, *lilliputian*. *Development* **128**, 801–813 (2001).
 89. von Dassow, G., Meir, E., Munro, E. M. & Odell, G. M. The segment polarity network is a robust developmental module. *Nature* **406**, 188–192 (2000).
 90. Basler, K. & Struhl, G. Compartment boundaries and the control of *Drosophila* limb pattern by *hedgehog* protein. *Nature* **368**, 208–214 (1994).
 91. Holleman, T., Bellefroid, E. & Pieler, T. The *Xenopus* homologue of the *Drosophila* gene *tailless* has a function in early eye development. *Development* **125**, 2425–2432 (1998).
 92. Stauber, M., Jäckle, H. & Schmidt-Ott, U. The anterior determinant *bicoid* of *Drosophila* is a derived Hox class 3 gene. *Proc. Natl Acad. Sci. USA* **96**, 3786–3789 (1999).
 93. Shaw, P. J., Salameh, A., McGregor, A. P., Bala, S. & Dover, G. A. Divergent structure and function of the *bicoid* gene in Muscoidea fly species. *Evol. Dev.* **3**, 251–262 (2001).
 94. McGregor, A. P. *et al.* Rapid restructuring of *bicoid*-dependent *hunchback* promoters within and between Dipteran species: implications for molecular co-evolution. *Evol. Dev.* **3**, 397–407 (2001).
 95. Dawes, R., Dawson, I., Falciani, F., Tear, G. & Akam, M. *Dax*, a locust Hox gene related to *fushi tarazu* but showing no pair-rule expression. *Development* **120**, 1561–1572 (1994).
 96. Davis, G. K., Jaramillo, C. A. & Patel, N. H. Pax group III genes and the evolution of insect pair-rule patterning. *Am. Zool.* **40**, 992 (2000).
 97. Schroeder, R., Jay, D. G. & Tautz, D. Elimination of Eve protein by CALI in the short germ band insect *Tribolium* suggests a conserved pair-rule function for even-skipped. *Mech. Dev.* **80**, 191–195 (1999).
 98. Salazar-Ciudad, I., Newman, S. A. & Solé, R. V. Phenotypic and dynamical transitions in model genetic networks I. Emergence of patterns and genotype-phenotype relationships. *Evol. Dev.* **3**, 84–94 (2001).
 99. Salazar-Ciudad, I., Solé, R. V. & Newman, S. A. Phenotypic and dynamical transitions in model genetic networks II. Application to the evolution of segmentation mechanisms. *Evol. Dev.* **3**, 95–103 (2001).

Acknowledgements

I thank P. Ahlberg and A. Panchen for comments on the manuscript; J. Blackburn, A. Cherrill and P. Griffin for photographs; P. Giblin for production of the typescript; and the Natural Environment Research Council for financial support for my current research.

Correspondence and requests for materials should be addressed to the author (e-mail: wallace.arthur@sunderland.ac.uk).

Crossover between classical and quantum shot noise in chaotic cavities

S. Oberholzer*, E. V. Sukhorukov*† & C. Schönenberger*

* Institut für Physik, Universität Basel, Klingelbergstrasse 82, 4056 Basel, Switzerland

† Present address: University of Florida, Gainesville, Florida 32611-8440, USA

The discreteness of charge in units of e led Schottky in 1918 to predict that the electrical current in a vacuum tube fluctuates even if all spurious noise sources are eliminated carefully¹. This phenomenon is now widely known as shot noise. In recent years, shot noise in mesoscopic conductors, where charge motion is quantum-coherent over distances comparable to the system size, has been studied extensively^{2–5}. In those experiments, charge does not propagate as an isolated entity through free space, as for vacuum tubes, but is part of a degenerate and quantum-coherent Fermi sea of charges. It has been predicted that shot noise in mesoscopic conductors can disappear altogether when the system is tuned to a regime where electron motion becomes classically chaotic⁶. Here we experimentally verify this prediction by using chaotic cavities where the time that electrons dwell inside can be tuned⁷. Shot noise is present for large dwell times, where the electron motion through the cavity is ‘smeared’ by quantum scattering, and it disappears for short dwell times, when the motion becomes classically deterministic.

Noise is generally due to randomness, which can be classical or quantum in nature. Randomness can be inherent in the reservoirs emitting charges or in the transmission process between emitter and collector. Shot noise observed in vacuum tubes^{8,9} (Fig. 1a) is solely due to fluctuations of the state occupancy in the reservoirs, leading to the completely random emission of electrons into vacuum. This random emission results in a spectral density S of the current fluctuations given by Schottky’s formula $S = 2e|I|$ (I denotes the mean electrical current). But the transmission from cathode (K in Fig. 1a) to anode (A) is free of randomness, because an emitted electron will eventually reach the anode with absolute certainty. Loosely speaking, the vacuum is noiseless.

In a coherent electrical conductor at zero temperature, the situation is the other way around (Fig. 1b). The emission from an ideal reservoir (a Fermi gas) into the wire is noiseless¹⁰. Here,

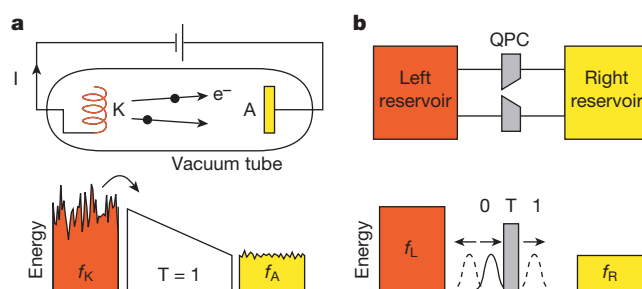


Figure 1 Origin of noise. **a**, In a classical vacuum tube, the randomness that gives rise to shot noise stems solely from fluctuations in the reservoirs. The transmission from cathode (K) to anode (A) through the vacuum is noiseless, because it occurs with unit probability. **b**, A quantum point contact (QPC)^{12,13} as a coherent conductor: at zero temperature randomness is due to scattering within the conductor only. Although this scattering can generally be classical or quantum-mechanical, only quantum scattering generates noise. (Here f denotes the distribution of electrons within the respective reservoirs and T the transmission probability.)

randomness is caused by scattering within the conductor, which can be either classical (deterministic) or quantum-mechanical. Both add to the electrical resistance, but only quantum scattering generates noise¹¹. This fundamental statement is verified experimentally in the present work by measuring the shot noise of open chaotic cavities.

Figure 2a shows a schematic drawing of a chaotic cavity connected via two narrow constrictions to a left (L) and a right (R) reservoir. At the contacts, charge transport takes place within an integer number N of energy channels, each with transmission probability $T = 1$. As electrons pass the contacts with unit probability, they are noiseless. Nevertheless, there is a fundamental resistance $R_{L,R}$ associated with each contact: $R_{L,R}^{-1} = G_{L,R} = (2e^2/h)N_{L,R}$ (refs 12 and 13). Inside the cavity the electron motion is chaotic. An electron entering the cavity from the left contact scatters inside the cavity and leaves it after the dwell time τ_D either at the left or the right contact with probabilities determined by the conductances $G_{L,R}$. The randomization inside the cavity leads to a total resistance R which is simply the sum of the left and right contact resistances: $R = R_L + R_R$ (the region inside the cavity has negligible resistance). Note that $R = R_L + R_R$ holds irrespective of whether chaos within the cavity is classical or quantum. As we will show below, this is markedly different for shot noise, the occurrence of which relies on the presence of quantum scattering—that is, diffraction caused by the wave nature of electrons. Because the degree of diffraction can be tuned in the experiment by changing the effective dwell time τ_D , this statement can experimentally be verified.

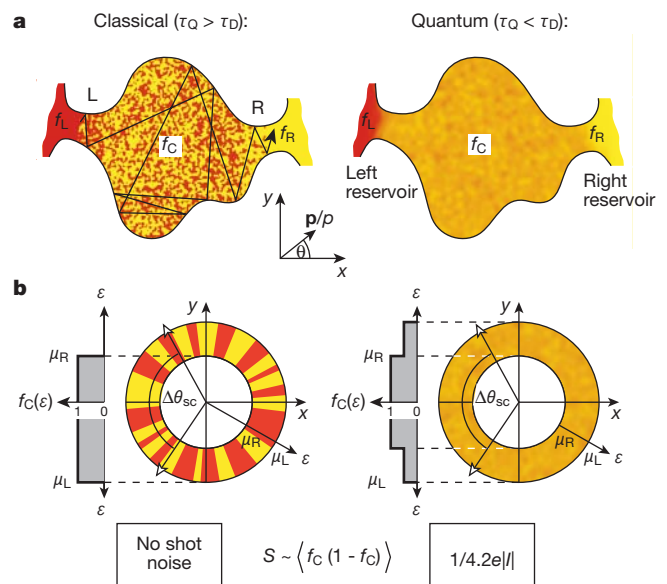


Figure 2 Chaotic cavity in the classical and quantum regime. **a**, Diagrams of a chaotic cavity connected by two contacts to a left (L) and right (R) reservoir. In the classical regime of deterministic scattering, the electrons with a specific momentum direction $\mathbf{p}/p$ inside the cavity either originate from the left (red) or the right (yellow) reservoir. The chaotic motion leads to a distribution function f_C inside the cavity that randomly switches between 0 (yellow) and 1 (red) within the relevant energy interval $[\mu_R, \mu_L]$. This is illustrated by the speckle pattern within the cavity. In the quantum case, the wavefunction of the electrons is spread over the whole cavity and we are not able to decide whether an electron stems from the left or the right side. **b**, The shot noise power is determined by the fluctuations $\langle f_C(1 - f_C) \rangle$ of the state occupancy f_C . Because classically f_C takes on either the value 0 or 1, the noise $S \propto \langle f_C(1 - f_C) \rangle$ is zero in this case, although $\langle f_C \rangle \neq 0, 1$. If quantum diffraction occurs, f_C can take on an arbitrary value between 0 and 1 within the energy interval $[\mu_R, \mu_L]$. In the ‘full’ quantum limit (strong diffraction) $\langle f_C(1 - f_C) \rangle = \langle f_C \rangle(1 - \langle f_C \rangle)$ leading to a shot noise of $(1/4)2e|I|$ for a symmetric cavity, because $\langle f_C \rangle = 1/2$ in this case.

The shot noise power of an open cavity can be expressed in terms of the distribution function f_c of the electrons inside the cavity¹⁴:

$$S = (2/R) \int d\epsilon \langle f_c(\mathbf{p}, \mathbf{r}) [1 - f_c(\mathbf{p}, \mathbf{r})] \rangle \quad (1)$$

Here $\langle \dots \rangle$ denotes the average over momentum direction $\mathbf{p}/p$ and spatial coordinate $\mathbf{r}$, and ϵ denotes the energy. There are three important timescales in the problem: (1) the ballistic flight time τ_F , which is the time an electron takes to traverse the cavity once; (2) the dwell time τ_D , and (3), the quantum scattering time τ_Q , which qualitatively is the mean time during which the classical trajectory is 'lost' by diffraction. Because of the relatively large cavities and small openings in the present experiments, $\tau_D \gg \tau_F$.

We first look at the classical regime, characterized by $\tau_Q \gg \tau_D$. In this regime classical (deterministic) trajectories are well defined. The distribution function f_c at a given point within the cavity and for a given energy ϵ can be expressed as a function of momentum direction $\mathbf{p}/p$ (Fig. 2). As the cavity is chaotic, a large number of different trajectories cross this given point. However, because classical physics is deterministic, each trajectory can be traced back to its unique origin, which is either the left or the right contact. Hence f_c is either one or zero. The product $f_c(1 - f_c)$, therefore, always equals zero and shot noise is absent. Intuitively, this appears to be very surprising, because of the presence of chaos. Classical chaos leads to a strange function f_c , which only takes on the values 0 and 1, but may switch between these two values in a very erratic and dense way (Fig. 2b). This, however, does not produce noise.

If, on the other hand, quantum diffraction cannot be neglected, that is, $\tau_Q < \tau_D$, the situation is very much changed. Owing to quantum scattering on impurities and at the boundary of the cavity, an electron wave (classically the 'trajectory') may split into two or

more partial waves leaving the cavity at different exits. A momentum state within the cavity cannot be traced back unambiguously to either contact, but carries information of both contacts simultaneously. Consequently, f_c is a weighted sum of f_L and f_R and can now take on values between 0 and 1. This uncertainty of not knowing where the electron came from and where it will go to is the source of noise.

The crossover between classical and quantum regimes in chaotic cavities has been theoretically discussed⁶. For a symmetric cavity, the noise power S is given by:

$$S = S_{eq} [1 + F(\beta \coth \beta - 1)], \quad \beta = \frac{eV}{2k_B \Theta} \quad (2)$$

where $F \equiv S/2e|I|$ is the Fano factor defined as the current-normalized noise power at zero temperature, and $S_{eq} = 4k_B \Theta/R$ the equilibrium (thermal) noise of the contacts. V denotes the applied voltage, and Θ the temperature. The Fano factor F decays to zero in the classical limit where $\tau_Q/\tau_D \rightarrow \infty$ and is to first order in τ_Q/τ_D given by:

$$F = \frac{1}{4} \left(1 - \frac{\tau_Q}{\tau_D} \right) \quad (3)$$

The noise power S can be written as $S = S_{CL} + S_Q$, where $S_{CL} = S_{eq}$. This part is purely classical. In contrast, the second non-equilibrium part $S_Q \propto F$ is a sensitive probe of quantum mechanics.

Cavities of size L comparable to the mean free path l are experimentally realized in a two-dimensional electron gas (Fig. 3a, right inset). Two quantum point contacts, electrostatically defined by metallic split gates, connect the cavity to the reservoirs. On the lateral two sides the electrons are confined by wet chemical etching of the two-dimensional electron gas. The openings of the cavity can be altered by varying the gate voltages. So-called 'open' cavities are defined when both quantum point contacts are adjusted

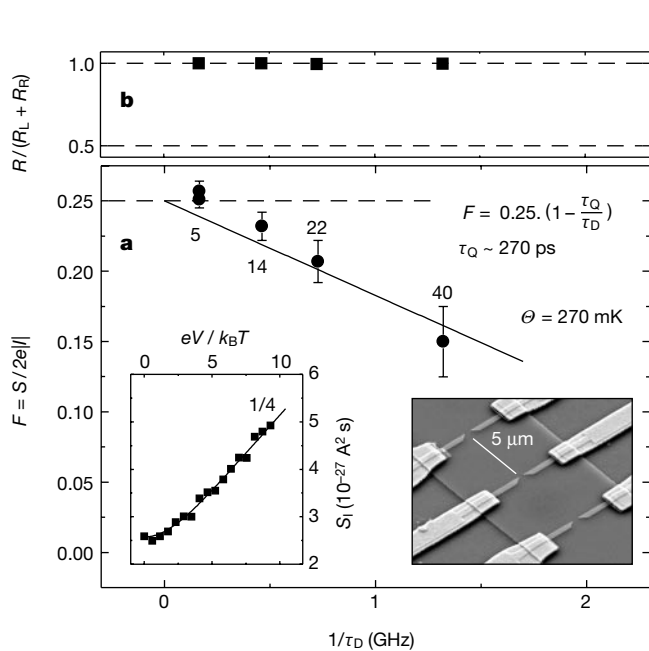


Figure 3 Fano factor $F \equiv S/2e|I|$ versus inverse dwell time τ_D^{-1} of a symmetric cavity ($N_L = N_R$). **a**, With increasing τ_D^{-1} , F becomes smaller than $1/4$ (quantum limit), indicating a crossover from full quantum to deterministic (classical) scattering inside the cavity. The numbers at the data points indicate the number of fully transmitted modes N at the contacts. **b**, Measured ratio between the total resistance R and the series resistance $R_L + R_R$ of the two contacts as a function of τ_D^{-1} . As the ratio $R/(R_L + R_R)$ is always 1, the electronic motion inside the cavity is chaotic. Left inset, shot noise for $N = 5$ in the full quantum regime, resulting in a Fano factor of $1/4$. Right inset, scanning electron microscope picture of the cavity.

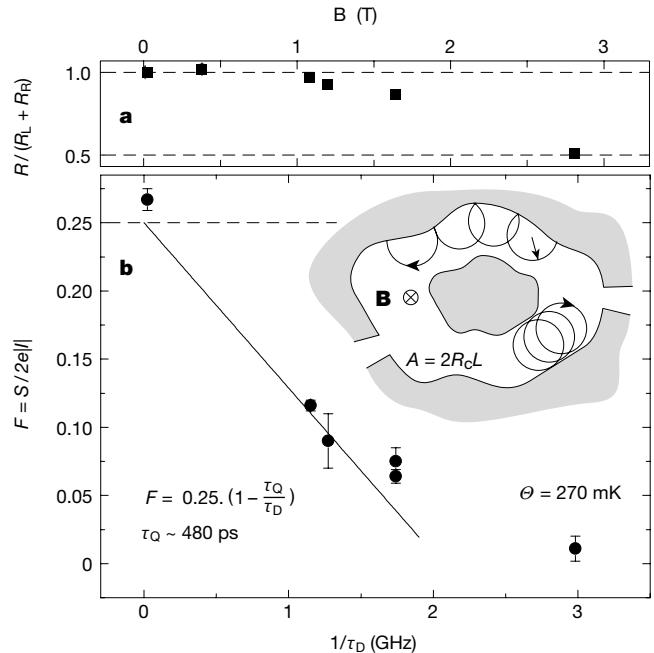


Figure 4 Measured Fano factors in various magnetic fields. A magnetic field reduces the dwell time: $\tau_D^{-1} \propto B$. Theoretically, this effect can be described as a change of the area of the cavity, because for fields corresponding to a cyclotron radius R_c smaller than the size L of the cavity, an annulus is formed at the edge with an inner part of the cavity not contributing to transport and noise (see inset). **a**, Total resistance R normalized to the series resistance $R_L + R_R$ as a function of the inverse dwell time τ_D^{-1} . **b**, F deviates from the quantum limit $1/4$ with increasing τ_D^{-1} , indicating a crossover from full quantum to classical scattering inside the cavity similar to the measurements in zero magnetic field (Fig. 3a).

to a conductance plateau with an integer number N of fully transmissive ($T = 1$) modes. Experimental details are given in ref. 7.

The deviations from the quantum limit¹⁵ $F = 1/4$ towards the regime of deterministic scattering $F < 1/4$, described by equation (3), are explored by changing τ_D . This dwell time depends on the area A of the cavity and the conductances $G_{L,R}$ of the contacts⁷:

$$\tau_D = \frac{2e^2 m}{\pi \hbar^2} \frac{A}{(G_L + G_R)} \quad (4)$$

In a first experiment we changed the openings (conductances) of the contacts to alter τ_D . In Fig. 3a, the Fano factor F of a symmetric ($N_L = N_R$) cavity is plotted as a function of the inverse dwell time τ_D^{-1} for four different settings ($N_L = N_R = 5, 14, 22$ and 40). The left inset shows the measured shot noise for $N_L = N_R = 5$. As the contacts were further opened and the dwell time was subsequently reduced, the shot noise was observed to decrease. The Fano factor F shows a pronounced decay below the quantum limit $1/4$. A linear fit of the data to equation (3) yields a τ_Q of 270 ps. We emphasize that the total resistance R equals the series resistance $R_L + R_R$ of the two contacts within the measurement accuracy of $\sim 3\%$ (Fig. 3b). Hence, the shot noise measurements are carried out in a regime where the direct (ballistic) transmission of electrons from the left to the right contact can be neglected. The suppression of shot noise observed here is a consequence of reduced diffraction, and serves to demonstrate that shot noise disappears in the limit of purely classical scattering.

An alternative way to change τ_D is to apply a perpendicular magnetic field. Because a magnetic field forces the electrons onto circular orbits with the cyclotron radius $R_c = mv_F/eB$ (where m denotes the electron mass and v_F the Fermi velocity), the dwell time τ_D will be reduced with increasing magnetic field B provided that $R_c < L$. An annulus of skipping orbits is formed (Fig. 4b inset) in which transport takes place. Such an annulus represents a 'new' (and smaller) chaotic cavity inside the actual cavity. For low magnetic fields (large filling factors), the electron dynamics inside the annulus can still be considered to be random (because of impurities or irregularities in the geometry of the cavity). Thus equations (2) and (3) are still valid with the area A in equation (4) replaced by $A = 2R_c L_c$, where $L_c \approx L$ is the circumference of the cavity. This leads to $\tau_D^{-1} \propto B$.

Figure 4b shows the measured Fano factor F as a function of inverse dwell time τ_D^{-1} in a magnetic field. We again observe a very marked reduction of F with increasing τ_D^{-1} , while the total resistance R approximately equals the series resistance $R_L + R_R$ of the two contacts in lower fields ($B < 1.2$ T) (Fig. 4a). A linear fit (equation (3)) results in a quantum scattering time τ_Q of ≈ 485 ps, which is in qualitative agreement with τ_Q obtained from the measurements in zero field. If the magnetic field is further increased beyond 1.2 T, the ratio $R/(R_L + R_R)$ starts to deviate from unity and equals $1/2$ at the highest magnetic field. Here, we enter a new regime in which a significant fraction of electrons is ballistically transmitted from source to drain. The last measurement point in Fig. 4 with $F = 0$ and $R/(R_L + R_R) \approx 0.5$ corresponds to the integer quantum Hall regime with filling factor four, where the electrons propagate within ballistic edge states. $\square$

Received 8 August; accepted 12 December 2001.

1. Schottky, W. Über spontane Stromschwankungen in verschiedenen Elektrizitätsleitern. *Ann. Phys. (Leipz.)* **57**, 541–568 (1918).
2. Blanter, Y. M. & Büttiker, M. Shot noise in mesoscopic conductors. *Phys. Rep.* **336**, 1–166 (2000).
3. Lesovik, G. B. Excess quantum noise in 2D ballistic point contacts. *JETP Lett.* **49**, 592–594 (1989).
4. Reznikov, M., Heiblum, M., Shtrikman, H. & Mahalu, D. Temporal correlation of electrons: Suppression of shot noise in a ballistic quantum point contact. *Phys. Rev. Lett.* **75**, 3340–3343 (1995).
5. Kumar, A., Saminadayar, L., Glattli, D. C., Jin, Y. & Etienne, B. Experimental test of the quantum shot noise reduction theory. *Phys. Rev. Lett.* **76**, 2778–2781 (1996).
6. Agam, O., Aleiner, I. & Larkin, A. Shot noise in chaotic systems: Classical to quantum crossover. *Phys. Rev. Lett.* **85**, 3153–3156 (2000).
7. Oberholzer, S. et al. Shot noise by quantum scattering in chaotic cavities. *Phys. Rev. Lett.* **86**, 2114–2117 (2001).

8. Hartmann, C. A. Über die Bestimmung des elektrischen Elementarquantums aus dem Schrotheffekt. *Ann. Phys. (Leipz.)* **65**, 51–78 (1921).
9. Hull, A. W. & Williams, N. H. Determination of elementary charge e from measurements of the shot-effect. *Phys. Rev.* **25**, 147–173 (1925).
10. Büttiker, M. Scattering theory of current and intensity noise correlations in conductors and wave guides. *Phys. Rev. B* **46**, 12485–12507 (1992).
11. Beenakker, C. W. J. & van Houten, H. Semiclassical theory of shot noise and its suppression in a conductor with deterministic scattering. *Phys. Rev. B* **43**, 12066–12069 (1992).
12. van Wees, B. J. et al. Quantized conductance of point contacts in a two-dimensional electron gas. *Phys. Rev. Lett.* **60**, 848–851 (1988).
13. Wharam, D. A. et al. One-dimensional transport and the quantisation of the ballistic resistance. *J. Phys. C* **21**, L209–L214 (1988).
14. Blanter, Y. M. & Sukhorukov, E. V. Semiclassical theory of conductance and noise in open chaotic cavities. *Phys. Rev. Lett.* **84**, 1280–1283 (2000).
15. Jalabert, R. A., Pichard, J.-L. & Beenakker, C. W. J. Universal quantum signatures of chaos in ballistic transport. *Europhys. Lett.* **27**, 255–2660 (1994).

Acknowledgements

We thank H. Büttiker and H. Thomas for discussions. This work was supported by the Swiss National Science Foundation.

Correspondence and requests for materials should be addressed to C.S.
(e-mail: Christian.Schoenenberger@unibas.ch).

Observation of stimulated emission by direct three-photon excitation

Guang S. He, Przemyslaw P. Markowicz, Tzu-Chau Lin & Paras N. Prasad

Institute for Lasers, Photonics and Biophotonics, University at Buffalo, The State University of New York, Buffalo, New York 14260-3000, USA

Multiphoton processes, predicted¹ theoretically in 1931, were for a long time considered to be mainly of academic interest. This view changed when it was shown^{2,3} that a two-photon absorption process could, because of a quadratic dependence of excitation on intensity, produce a spatially confined excitation useful for three-dimensional data storage and imaging. Two-photon absorption has received considerable attention recently because of the development of highly efficient two-photon-sensitive materials, leading to numerous technological applications^{4–28}. These successes have created interest in exploring applications based on three-photon excitations²⁹. For a three-photon process, a longer excitation wavelength such as those common in optical communications can be used. Also, the cubic dependence of the three-photon process on the input light intensity provides a stronger spatial confinement, so that a higher contrast in imaging can be obtained. Here we report the observation of a highly directional and up-converted stimulated emission as an amplified spontaneous emission, produced in an organic chromophore solution by a strong simultaneous three-photon absorption at 1.3 μm . This achievement suggests opportunities for a three-photon process in frequency-upconversion lasing, short-pulse optical communications, and the emerging field of biophotonics.

In our experiment, the gain medium is the organic chromophore 4-[N-(2-hydroxyethyl)-N-(methylamino phenyl)-4'-(6-hydroxyhexyl sulphonyl)] stilbene (APSS) dissolved in dimethyl sulfoxide (DMSO). APSS has been reported¹² to exhibit good properties for two-photon pumped lasing as well as for optical power limiting, when excited at a wavelength of ~ 800 nm. The molecular structure of APSS and its linear absorption spectrum in DMSO were shown in our previous publications¹². We report here that APSS also exhibits a strong three-photon absorption when excited by ultrashort laser pulses at $\sim 1.3 \mu\text{m}$, producing population inversion. This creates the prospect of developing new three-photon materials with even more enhanced cross-sections for three-photon absorption. Our

study shows that when a focused ultrashort-pulse laser beam at a wavelength of $\sim 1.3\ \mu\text{m}$ is passed through a solution of APSS in DMSO, a yellowish-green and highly directional coherent emission is produced in both the forward and backward directions, provided that the pump intensity is higher than a certain threshold value. The visible emission wavelength is shorter than $1/2$, and longer than $1/3$, of the pump wavelength.

In our experiment, a 1-cm-long quartz cuvette filled with an APSS/DMSO solution of concentration $d_0 = 0.06\ \text{mol l}^{-1}$ was employed as the gain medium. A laser beam of high peak power and $\sim 1.3\text{-}\mu\text{m}$ wavelength was focused in the centre of the sample cuvette via a lens of focal length $f = 10\ \text{cm}$. This laser beam was obtained from an optical parametric generator pumped by a mode-locked Ti:sapphire laser oscillator–amplifier system producing $\sim 150\text{-fs}$ -duration, $\sim 0.79\text{-}\mu\text{m}$ -wavelength laser output, with a spectral width of $\sim 8\ \text{nm}$, and at a repetition rate of $1\ \text{kHz}$. The APSS molecules in solution have no linear absorption in the spectral range from $\sim 500\ \text{nm}$ to $\geq 1.3\ \mu\text{m}$. But when illuminated by the focused $\sim 1.3\text{-}\mu\text{m}$ laser beam, an obvious yellowish-green fluorescence can be readily observed from any direction, which implies a frequency-upconversion emission process, induced by three-photon absorption. Furthermore, once the intensity level of the input laser beam is increased above a certain threshold value, a highly directional coherent emission of $\sim 0.55\ \mu\text{m}$ can be observed in both the forward and the backward directions. Figure 1a shows the intense and highly directional forward stimulated emission of $0.55\text{-}\mu\text{m}$ wavelength from the APSS solution cell; Fig. 1b shows second-harmonic generation (SHG) when the same $1.3\text{-}\mu\text{m}$ pump beam passes separately through a beta barium borate (BBO) crystal.

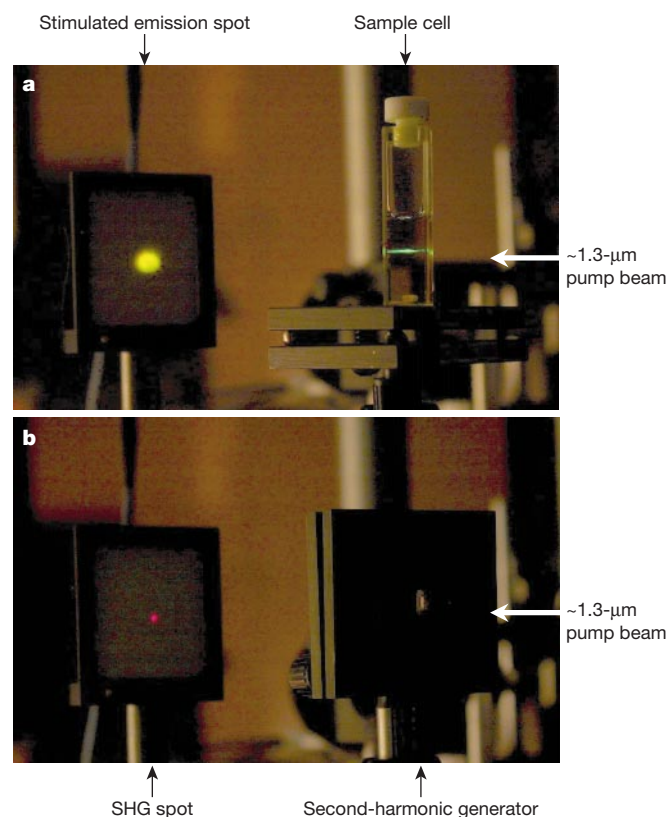


Figure 1 Photographs showing highly directional, frequency-upconverted stimulated emission, pumped by three-photon absorption. **a**, Forward stimulated emission (green) from an APSS solution cuvette pumped with a focused $1.3\text{-}\mu\text{m}$ beam; **b**, second-harmonic generation (SHG) produced separately from a BBO crystal excited with the same $1.3\text{-}\mu\text{m}$ pump beam.

These two are shown together for comparison of their beam coherence properties. The spectral and temporal properties of this coherent three-photon excited emission in both forward and backward directions are basically identical, and not dependent on the incident angle (from 0° to $\geq 10^\circ$) of the pump beam on the sample cuvette. These facts exclude the possibility that the observed directional emission is due to continuum generation, because the latter is not expected in the backward direction.

Figure 2 shows the spectral structures of the three-photon-induced spontaneous fluorescence measured at $\sim 90^\circ$ direction with respect to the pump beam, as well as of the forward and backward frequency-upconverted stimulated emission. It also shows the spectral characteristics of the second harmonic of the pump laser beam, which was generated by an external BBO crystal and could be used to determine the exact pump wavelength. From this figure we can see that the central wavelengths of the coherent visible emission in both directions are nearly the same, and located at $\sim 553\ \text{nm}$ with a spectral width of $\sim 10\ \text{nm}$ (full-widths at half-minimum) that is about 8.5 times narrower than the corresponding fluorescence spectrum. From the peak position of the SHG signal of the pump beam, the central wavelength of the pump light is determined to be $1,292\ \text{nm}$.

Figure 3a–c shows temporal profiles of the pump laser pulse, the forward up-conversion coherent emission, and the $\sim 90^\circ$ fluorescence emission, respectively. These profiles were measured utilizing a high-speed streak camera (C5680-22 from Hamamatsu) with a resolution of $\sim 2\ \text{ps}$ for Fig. 3a and b, and $\sim 20\ \text{ps}$ for Fig. 3c. Two features can be seen from this figure. First, there is a delay between the input pump pulse and the stimulated emission pulse, and

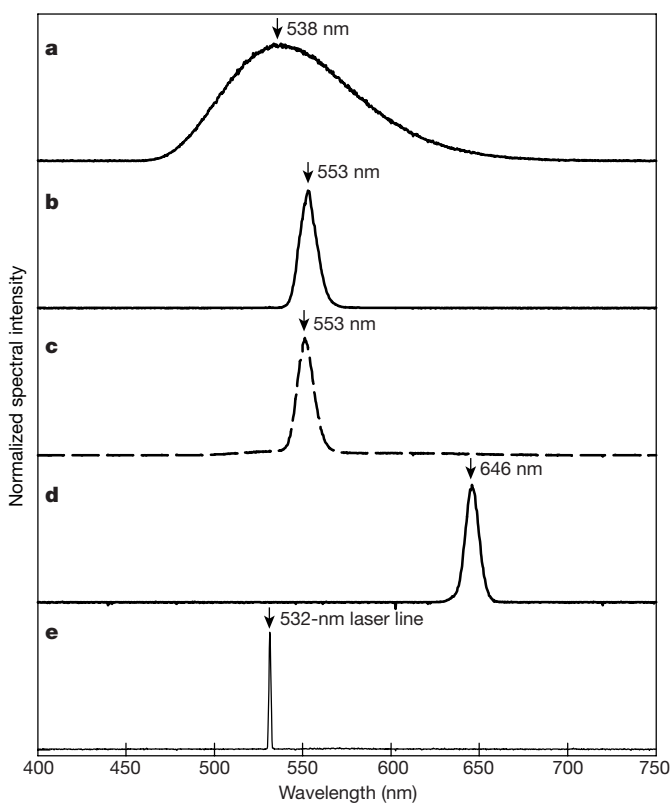


Figure 2 Spectra of the stimulated emission and induced fluorescence, pumped by three-photon absorption. **a**, Lateral fluorescence; **b**, forward stimulated emission; **c**, backward stimulated emission; **d**, second-harmonic generation of $\sim 1.3\text{-}\mu\text{m}$ pump radiation passing through a BBO crystal; and **e**, 532-nm laser line for calibration purpose showing the spectral resolution of the spectrometer used.

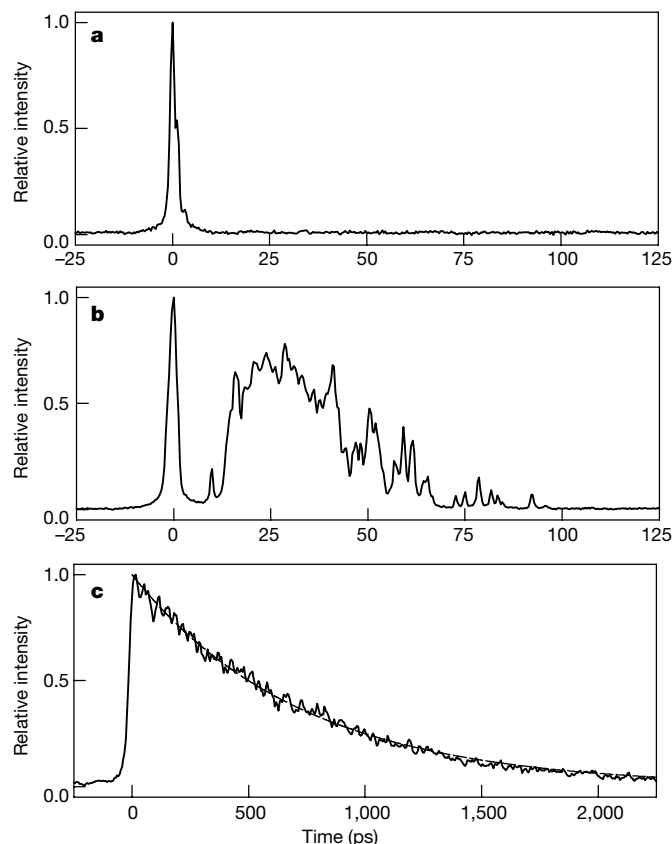


Figure 3 Streak-camera measurements. Temporal profiles of the 1.3- μm pump pulse (a), the transmitted pump pulse together with the 553-nm forward stimulated emission (b), and the three-photon-induced fluorescence emission (c). In c, the lifetime of the fluorescence is ~ 720 ps.

second, the pulse duration of stimulated emission is much longer (about 30–50 ps) than the pump pulse. The delay in emission may reflect an equivalent relaxation time of APSS molecules from their higher (three-photon) excited states to the emitting lower relaxed state. From the observed temporal behaviour of the visible coherent emission, the possibility of coherent emission from stimulated scattering and four-wave-mixing processes can be excluded, as these processes would not produce any overall delay or temporally broadened emission with respect to the pump pulse. As shown in Fig. 3c, the average fluorescence lifetime, determined from the side

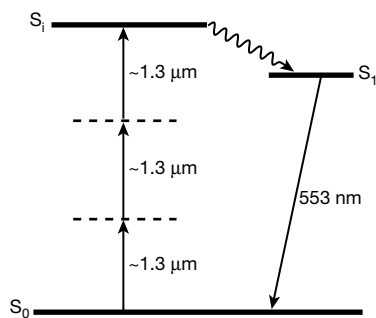


Figure 4 Diagram showing the proposed mechanism for three-photon excitation and stimulated emission. S_1 is a higher singlet (electronic or vibronic) state; S_1 is the radiative lower singlet state; S_0 is the ground state. Under our experimental conditions, a transient population inversion is created between S_1 and S_0 .

emission of the cuvette, is ~ 720 ps. The forward coherent emission, therefore, shows a considerable temporal narrowing, a gain characteristic of stimulated emission. Cavity feedback is not possible in our case because of the use of ultrashort pump pulse, so what we observed is a highly directional amplified spontaneous emission.

The suggested energy level structure and the simultaneous three-photon absorption pathway for our chromophore are shown in Fig. 4. The population inversion is created between the lower emitting state, S_1 , and the ground state, S_0 . To demonstrate the high directionality of the stimulated emission, the transmitted pump laser beam and the forward stimulated emission beam were recollimated through a lens, and then projected on a paper screen. Images of these two beams on the screen could be recorded separately by a charge-coupled device (CCD) camera in conjunction with appropriate filters. We find that these two beams (also including the backward coherent emission beam) have nearly the same spatial profiles.

On the basis of these results—that is, the remarkable spectral narrowing, high directionality, temporal delay, and a relatively longer emission pulse compared to the pump pulse—we conclude that the observed coherent visible emission is a single-pass stimulated emission in the form of amplified spontaneous emission

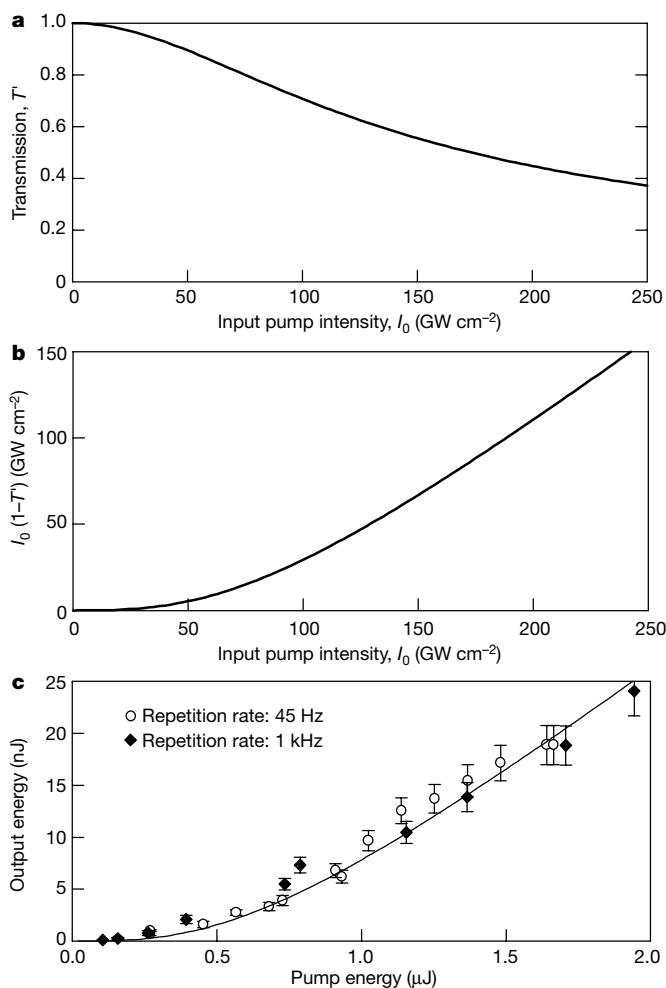


Figure 5 Output/input characteristics of three-photon-pumped stimulated emission. a, Transmission change induced by three-photon absorption versus input intensity of 1.3- μm radiation. $\gamma = 0.5 \times 10^{-4} \text{ cm}^3 \text{ GW}^{-2}$, $z = 1 \text{ cm}$ (see text for definitions). b, Pump intensity depletion versus input intensity. c, Output optical energy versus the pump energy.

derived from population inversion of the excited chromophore molecules, produced by a simultaneous three-photon absorption process. The nonlinear transmissivity of a three-photon absorbing medium can be written as²⁹

$$T' = \frac{I(z)}{I_0} = \frac{1}{\sqrt{(1 + 2\gamma z I_0^2)}} \quad (1)$$

where I_0 is the input light intensity, z is the propagation distance in the medium, and γ is the three-photon absorption coefficient of the medium.

From equation (1), the γ value for a given sample medium can be experimentally determined by measuring the nonlinear transmission at a given input intensity level. In our case, at a pump level of $\sim 190 \text{ GW cm}^{-2}$, the γ value is evaluated to be $0.5 (\pm 15\%) \times 10^{-4} \text{ cm}^3 \text{ GW}^{-2}$ for the APSS/DMSO solution of $d_0 = 0.06 \text{ mol l}^{-1}$. On the basis of this value, we can predict the theoretical curve of nonlinear transmission T' of the pump beam versus its intensity change (Fig. 5a). The depletion of pump intensity (or energy) due to three-photon absorption is determined by $I_0(1 - T')$, and is shown in Fig. 5b. It may be assumed that the stimulated emission is approximately proportional to the absorbed pump energy. The measured values of pulse energy of the overall (forward and backward) visible stimulated emission as a function of the input pump energy are shown in Fig. 5c. These experimental data were obtained at two different repetition rates for the pump pulses, while a beam-chopper was employed in the low-repetition-rate case. In Fig. 5c, the solid fitting curve is obtained directly from Fig. 5b after a simple re-scaling of the y axis. The agreement between the experimental data and the fitting curve seems to support the three-photon absorption model. As shown in Fig. 5c, at the pump level of $\sim 1.5 \mu\text{J}$, the output optical energy was $\sim 17 \text{ nJ}$, with an overall energy conversion coefficient of $\sim 1.1\%$. At this pump level, the measured nonlinear transmission was ~ 0.47 ; therefore, the net conversion coefficient from the absorbed pump energy to the output energy would be $\sim 2.1\%$.

Efficient three-photon absorption at $1.3 \mu\text{m}$ could be useful for short-pulse optical fibre communications, as it could be used for wavelength shifting, pulse reshaping and stabilization. Because the $1.3\text{-}\mu\text{m}$ radiation also provides a greater penetration into biological tissues and is less damaging to the cells than the visible-wavelength radiation, an efficient three-photon absorption at $1.3 \mu\text{m}$ as reported here could also provide opportunities in biophotonics, such as bioimaging and light-activated therapy at deep levels in tissues. □

Received 19 September; accepted 20 December 2001.

- Göppert-Meyer, M. Elementartakte mit zwei Quantensprüngen. *Ann. Phys.* **9**, 275–294 (1931).
- Parthenopoulos, D. A. & Rentzepis, P. M. Three-dimensional optical storage memory. *Science* **245**, 843–845 (1989).
- Denk, W., Strickler, J. H. & Webb, W. W. Two-photon laser scanning fluorescence microscopy. *Science* **248**, 73–76 (1990).
- Zhao, C. F., He, G. S., Bhawalkar, J. D., Park, C. K. & Prasad, P. N. Newly synthesized dyes and their polymer/glass composites for one- and two-photon pumped solid-state cavity lasing. *Chem. Mater.* **7**, 1979–1983 (1995).
- Reinhardt, B. A. *et al.* Highly active two-photon dyes: design, synthesis, and characterization toward application. *Chem. Mater.* **10**, 1863–1874 (1998).
- Kim, O.-K. *et al.* New class of two-photon-absorbing chromophores based on dithienothiophene. *Chem. Mater.* **12**, 284–286 (2000).
- Andronov, A. *et al.* Novel two-photon absorbing dendritic structures. *Chem. Mater.* **12**, 2838–2841 (2000).
- Kannan, R. *et al.* Diphenylaminofluorene-based two-photon-absorbing chromophores with various π -electron acceptors. *Chem. Mater.* **13**, 1896–1904 (2001).
- Ehrlich, J. B. *et al.* Two-photon absorption and broadband optical limiting with bis-donor stilbene. *Opt. Lett.* **22**, 1843–1845 (1997).
- Albota, M. *et al.* Design of organic molecules with large two-photon absorption cross sections. *Science* **281**, 1653–1656 (1998).
- Belfield, K. D., Schafer, K. J., Mourad, W. & Reinhardt, B. A. Synthesis of new two-photon absorbing fluorene derivatives via Cu-mediated Ullmann condensations. *J. Org. Chem.* **65**, 4475–4481 (2000).
- He, G. S., Bhawalkar, J. D., Zhao, C. F., Park, C.-K. & Prasad, P. N. Two-photon-pumped cavity lasing in a dye-solution-filled hollow-fiber system. *Opt. Lett.* **20**, 2393–2395 (1995).
- He, G. S., Yuan, L., Cui, Y., Li, M. & Prasad, P. N. Studies of two-photon pumped frequency-upconverted lasing properties of a new dye material. *J. Appl. Phys.* **81**, 2529–2537 (1997).

- He, G. S., Signorini, R. & Prasad, P. N. Longitudinally two-photon pumped leaky waveguide dye film laser. *IEEE J. Quantum Electron.* **34**, 7–13 (1998).
- He, G. S. *et al.* Two-photon absorption and optical-limiting properties of novel organic compounds. *Opt. Lett.* **20**, 435–437 (1995).
- He, G. S. *et al.* Nonlinear optical properties of a new chromophore. *J. Opt. Soc. Am. B* **14**, 1079–1087 (1997).
- He, G. S., Weder, C., Smith, P. & Prasad, P. N. Optical power limiting and stabilization based on a novel polymer compound. *IEEE J. Quantum Electron.* **34**, 2279–2285 (1998).
- Dvornikov, A. S. & Rentzepis, P. M. Novel organic ROM materials for optical 3D memory devices. *Opt. Commun.* **136**, 1–6 (1997).
- Strickler, J. H. & Webb, W. W. Three-dimensional optical data storage in refractive media by two-photon point excitation. *Opt. Lett.* **16**, 1780–1782 (1991).
- Pudavar, H. E., Joshi, M. P., Prasad, P. N. & Reinhardt, B. A. High-density three-dimensional optical data storage in a stacked compact disk format with two-photon writing and single photon readout. *Appl. Phys. Lett.* **74**, 1338–1340 (1999).
- Parthenopoulos, D. A. & Rentzepis, P. M. Two-photon volume information storage in doped polymer systems. *J. Appl. Phys.* **68**, 5814–5818 (1990).
- Gura, T. Biologists get up close and personal with live cells. *Science* **276**, 1988–1990 (1997).
- Bhawalkar, J. D. *et al.* Nondestructive evaluation of polymeric paints and coatings using two-photon laser scanning confocal microscopy. *Polymer* **38**, 4551–4555 (1997).
- Wang, X. *et al.* A chemically labeled cytotoxic agent: Two-photon fluorophore for optical tracking of cellular pathway in chemotherapy. *Proc. Natl Acad. Sci. USA* **96**, 11081–11084 (1999).
- Bhawalkar, J. D., Kumar, N. D., Zhao, C. F. & Prasad, P. N. Two-photon photodynamic therapy. *J. Clin. Laser Med. Surg.* **15**, 201–204 (1997).
- Joshi, M. P., Pudavar, H. E., Swiatkiewicz, J., Prasad, P. N. & Reinhardt, B. A. Three-dimensional optical circuitry using two-photon-assisted polymerization. *Appl. Phys. Lett.* **74**, 170–172 (1999).
- Cumpston, B. H. *et al.* Two-photon polymerization initiators for three-dimensional optical data storage and microfabrication. *Nature* **398**, 51–54 (1999).
- Kawata, S., Sun, H.-B., Tanaka, T. & Takada, K. Finer features for functional microdevices. *Nature* **412**, 697–698 (2001).
- He, G. S., Bhawalkar, J. D. & Prasad, P. N. Three-photon-absorption-induced fluorescence and optical limiting effects in an organic compound. *Opt. Lett.* **20**, 1524–1526 (1995).

Acknowledgements

This work was supported by the US Air Force Office of Scientific Research, Washington DC, and the Polymer Branch of the US Air Force Research Laboratory at Dayton.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to P.N.P. (e-mail: pnprasad@acsu.buffalo.edu).

Why stainless steel corrodes

Mary P. Ryan*, David E. Williams†, Richard J. Chater*, Bernie M. Hutton† & David S. McPhail*

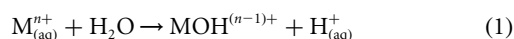
* Department of Materials, Imperial College of Science, Technology and Medicine, Prince Consort Road, London SW7 2BP, UK

† Department of Chemistry, University College London, 20 Gordon Street, London WC1H 0AJ, UK

Stainless steels are used in countless diverse applications for their corrosion resistance. Although they have extremely good general resistance, they are nevertheless susceptible to pitting corrosion. This localized dissolution of an oxide-covered metal in specific aggressive environments is one of the most common and catastrophic causes of failure of metallic structures. The pitting process has been described as random, sporadic and stochastic and the prediction of the time and location of events remains extremely difficult¹. Many contested models of pitting corrosion exist, but one undisputed aspect is that manganese sulphide inclusions play a critical role. Indeed, the vast majority of pitting events are found to occur at, or adjacent to, such second-phase particles^{2,3}. Chemical changes in and around sulphide inclusions have been postulated⁴ as a mechanism for pit initiation but such variations have never been measured. Here we use nanometre-scale secondary ion mass spectroscopy to demonstrate a

significant reduction in the Cr:Fe ratio of the steel matrix around MnS particles. These chromium-depleted zones are susceptible to high-rate dissolution that 'triggers' pitting. The implications of these results are that materials processing conditions control the likelihood of corrosion failures, and these data provide a basis for optimizing such conditions.

It is known that pitting corrosion propagates because an extreme solution condition is developed and maintained within a localized zone⁵. The contentious issue is how this condition is established in the first place. Stainless steel typically shows pitting corrosion in a neutral aqueous solution containing as little as 1 mM chloride, wherein the passive current density of dissolution of the steel is usually on the scale of nanoamperes per square centimetre. In contrast, within a propagating pit, we observe $\text{pH} < 0$, chloride concentration, $[\text{chloride}] > 1 \text{ M}$ and the dissolution current density is typically $> 1 \text{ A cm}^{-2}$. The stabilization of propagation of pitting corrosion is well understood. The corrosion current density within the developing pit is very high and this draws Cl^- anions into the pit by electromigration (they maintain charge neutrality). The local environment is acidified by the hydrolysis of the dissolving metal (M) cations



Thus the local environment is significantly more aggressive than the bulk solution and the metal inside the pit is maintained in an

actively dissolving state. There is a critical current density required to maintain this local solution in the face of diffusion of species out of the pit cavity. Occlusion of the pit cavity by precipitated corrosion products or partly dissolved lace-like metallic covers serves to restrict diffusion and maintain stability.

To understand the initiation process we need to explain how the extreme conditions arise that are needed to dissolve the steel. All current models concerning this point are speculative. Several have been proposed that rely on properties of the protective oxide films and are generally independent of inclusions^{6–10}. However, the practical evidence for stainless steels is that pits occur at sulphide inclusions that are typically enriched in MnS (but not at all at other precipitates, such as oxides) so a model that does not address the effect of sulphide inclusions will not provide an adequate answer to the problem. Extensive corrosion tests documented in the literature have shown that local electrochemical activity is associated with the presence of sulphide inclusions. In particular, Suter and Bohni¹¹ have used micro-electrochemical techniques to show that current transients associated with metal dissolution, where the current typically rises within 1–100 s by 0.1–10 μA before dropping abruptly to the baseline value, are greater adjacent to such inclusions in stainless steel than at any other locations on the metal surface. Ideas on pit initiation that explicitly include inclusion effects have tended to focus on the role of soluble dissolution products. The formation of sulphates², elemental sulphur³ or

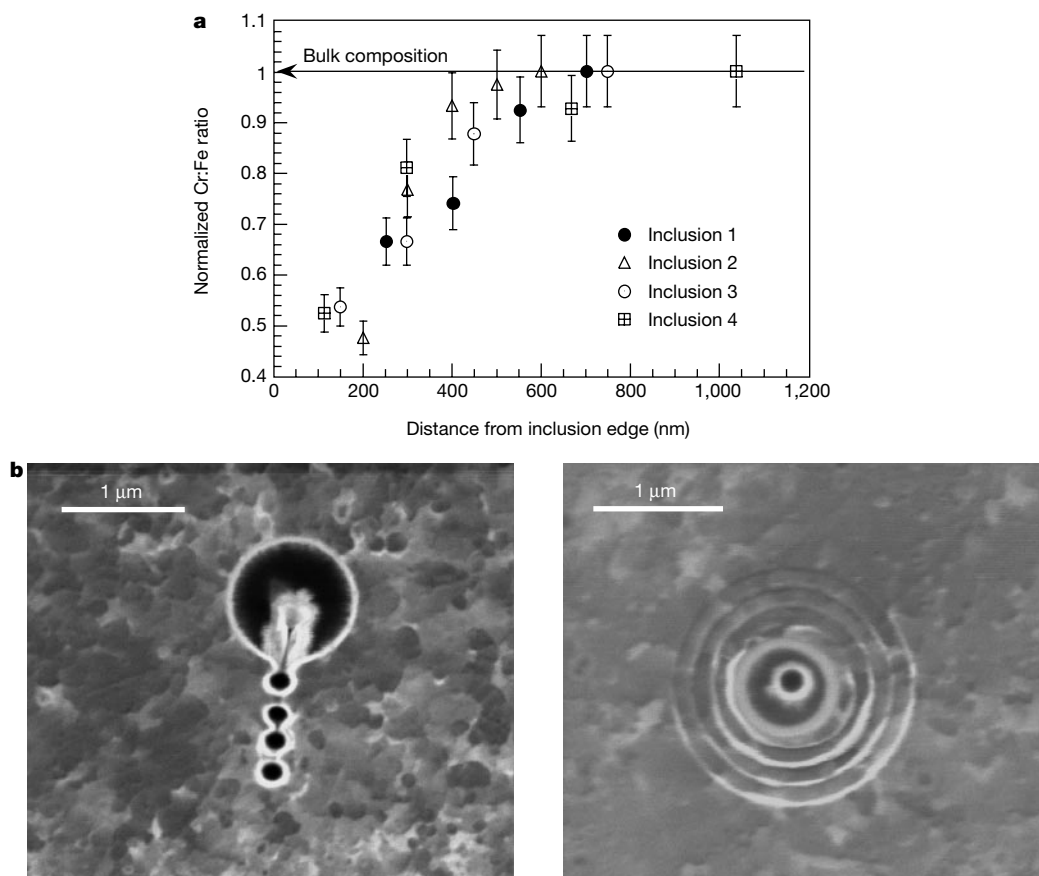


Figure 1 Local FIB-SIMS analysis adjacent to MnS particles in 316 stainless steel. **a**, Normalized Cr:Fe ratio as a function of proximity to a sulphide inclusion. Four different inclusion sites on the surface are shown and the same trend was evident at many other sites. The data were obtained using an FEI focused ion beam system with SIMS analysis to collect mass spectra at defined locations on the sample surface. A gallium beam was used with a beam width of 100 nm, a current of 100 pA and energy of 30 keV. The peaks in the

mass spectra were integrated at fixed width ($M \pm 0.4$ where M is the mass number of the element of interest) and no other fitting parameters were used. **b**, Secondary electron images of the surfaces after SIMS analysis. The beam shape was either a spot (left) or annulus (right). The inclusion is the large dark centre circle; the smaller features correspond to sputter craters where the SIMS data were collected. Images were obtained with a beam current of 10 pA and energy 30 keV. Scale bars, 1 μm .

Table 1 Analysis of 316F stainless steel

	Fe	Al	Ni	Si	Cu	Ti	Mn	As	S	P
Analysis		0.10	8.61	0.51			0.67	0.02	0.16	0.02
Electron probe	69.4 ± 1.3	ND	8.33 ± 0.28	ND	0.01 ± 0.01	0.002	0.70 ± 0.05	ND	0.05 ± 0.01	0.03 ± 0.02
	Cr	Mo	Co	Sn	V	Sr	Pb	Nb	C	B
Analysis	17.65	2.19	0.035	0.005		3.55	0.0014	0.029	0.062	0.0024
Electron probe	17.89 ± 0.24	2.02 ± 0.06	0.083 ± 0.012	0.003	0.023 ± 0.007					

Analysis is from the Bureau of Analysed Standards, Cleveland, UK. Electron probe microanalysis results (ND, not detectable; blank entry, not determined) are an average of 31 points in the metal (mean ± standard deviation), and exclude inclusions. The standard deviation in Fe, Cr, Ni was largely due to a large-scale, correlated variation with a pattern suggesting an association with the cooling or rolling of the steel. Sn, V, Cu were very variable. Al, Si, S, As and Ti were concentrated in inclusions. Results expressed as weight per cent.

thiosulphate¹² by oxidation of the inclusion, or of H₂S by simple chemical dissolution², have all been suggested as reasons for the deleterious effect. An alternative suggestion has been that elemental sulphur can enrich at the metal surface where it lowers the activation energy for dissolution and inhibits repassivation (reformation of the protective oxide)¹³. However, the problem with these models is that, in the dilute, neutral environments in which pitting corrosion can begin, the sulphides which constitute the inclusions are themselves passive.

More recent work shows that pitting corrosion is triggered by an unusual, high-rate dissolution of inclusions⁴. These workers⁴ propose that pitting has its genesis in the fabrication of the steel itself: when the steel is poured, the metal solidifies before the sulphide inclusions do, and as it cools, the composition of the inclusion itself and of the metal zone around it will change. Although the critical thermodynamic data are either imprecise or unavailable, they⁴ indicated that the replacement of Mn by Cr and Fe within the sulphide would occur. This has two consequences in terms of corrosion: (1) the inclusion itself is less stable to dissolution owing to the presence of Fe-rich sulphide species and of Fe- and Mn-enriched metallic phases within the inclusion; and (2) around the inclusion the alloy matrix has a chromium-depleted zone. Up until now there has been no data to support such a model, even

though it is certainly testable. This is primarily because of the difficulty of making analytical measurements with the required resolution. Scanning transmission electron microscopy (STEM) measurements would have the capacity but would rely on making a section through an active inclusion (the inclusion itself is only 1 µm in diameter). Scanning Auger measurements have been reported but only after corrosion has occurred (and the inclusion may not be stable in the electron beam)¹⁴. We have now succeeded in using focused ion beam/secondary ion mass spectroscopy (FIB/SIMS) to provide characterization of the material chemistry as a function of the proximity to sulphide inclusion to achieve unprecedented resolution in such a study.

The specimen was type 316F stainless steel (a high-sulphur steel with a rich inclusion content; a sample with certified analysis was used; see Table 1) polished finally with 0.05-µm Al₂O₃ or 0.25-µm diamond, with careful ultrasonic cleaning in ethanol between each stage of polishing. A high-sulphur steel was chosen as this would provide a large number of MnS inclusions for analysis. While it is known that low-sulphur steels have improved resistance, localized corrosion in these steels is still always associated with MnS inclusions. In fact, in steels with this Cr content that are completely free of inclusions pitting does not occur¹³. Sulphide inclusions are inactive as sites for stable corrosion initiation if they are sufficiently

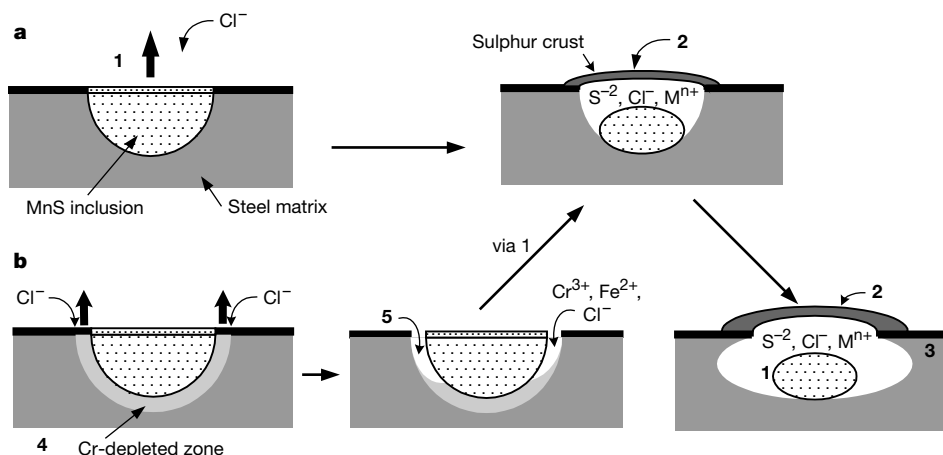


Figure 2 Schematic illustration of the process triggering pitting corrosion of stainless steel. **a**, Unanticipated high-rate electrochemical dissolution of (Mn,Cr)S inclusions (1) has been shown²⁰ to give rise to a sulphur-capped occluded zone (2) within which a sulphide- and chloride-rich acidic solution could develop through further dissolution of the inclusion. 'Stainless' steel has been shown to be unstable in contact with such a solution²¹. The parent metal therefore dissolves and the pit develops by undercutting the metal surface (3). The problem with such a scenario is that under normal conditions the inclusion itself should be stable—therefore, we need to consider what causes the rapid dissolution process (1) in the first place. **b**, Here we show Cr depletion (4) of the parent metal around the inclusion that can be related to the thermodynamics of interaction between sulphide and metal during casting of the steel⁴. We propose that pits trigger by dissolution (5) of the

chromium-depleted zone, creating an acidic environment within which the inclusion is unstable. Webb *et al.*¹⁹ have shown that narrow trenches (typical width 100 nm, commensurate with our measurement of chromium-depletion layer width) indeed form at the edge of sulphide inclusions and that pits trigger from these. Their modelling predicts rapid development of an extremely aggressive environment within these trenches if the steel dissolution current is at values commensurate with those found for steels depleted in chromium to the extent that we have observed at the inclusion edge. Hence the high rate of inclusion dissolution, which triggers the final breakdown, is itself started. Acidification of the cavities is by hydrolysis of CrCl₆³⁻ formed by dissolution either of the steel or of Cr₂S₃. Chloride enrichment occurs as a result of electromigration, to balance charge and carry current.

small. This is because a small open cavity formed by the initial triggering cannot retain a solution of composition that is sufficiently aggressive to continue the corrosion¹⁶. Other than this, there is no evidence that the mechanism of corrosion initiation at small inclusions is any different from that at large ones.

Analysis was performed using an FEI model FIB 200 focused ion beam with SIMS attachment. A primary gallium beam was used throughout the experiments (beam current, size and voltage are described in each figure legend).

Figure 1a shows the variation in the Cr:Fe ratio as a function of distance from the edge of an inclusion—this trend was observed at several inclusion locations on the sample surface. The depletion of chromium from the alloy matrix as the MnS particle is approached is clearly evident. The adjacent micrograph indicates the locations where SIMS analysis was performed, as indicated by the sputtered craters. The amount of depletion of Cr in the near-inclusion zone can be estimated from the variation from the bulk composition, which is known (~18%, Table 1). In these measurements we estimate that the alloy next to the inclusion has a Cr content of as little as 10 atomic per cent (assuming a constant ion yield of Fe and Cr in the Fe–xCr alloy matrix). In making this estimation we assume that any variations in the sputter rate, charging effects or yield enhancement owing to the presence of the inclusion will affect both the Fe and Cr signals in the same way and so will not affect the Fe/Cr ratio. To our knowledge, there are no data on ion yields for this system and we have assumed for simplicity that they remain constant. We will conduct a systematic study of ion and

sputter yields over a range of compositions of interest in future experiments.

The critical issue now becomes how much chromium depletion is required to render the steel unstable. For binary FeCr alloys with no second-phase particles, pitting was never observed above 16 at.% Cr (ref. 15). It has long been known that there is a critical amount of Cr (~13%) required to make 'stainless' steel and that below this threshold the material behaves much like iron¹⁷. Percolation models of the dissolution of binary alloys have been used to understand this behaviour and its consequences¹⁸. This suggests, therefore, that a reduction in the chromium content of only about 3.5% would be sufficient to make the material susceptible—and we find a depletion of this order, or significantly larger, around sulphide inclusions from our analysis. We suggest that the correct sequence of events for pitting corrosion is therefore as follows: (1) composition changes are induced in and around sulphide inclusions during the processing of the steel from the melting temperature of the metal to the solidification temperature of the inclusion; (2) the chromium-depleted zones provide rapid, high-rate metal dissolution events that cause a narrow trench to form at the inclusion edge and that change the local environment (by equation (1) above)¹⁹; (3) in this new local environment the inclusion itself is unstable to dissolution; (4) sulphur products form a crust around the former inclusion²⁰ providing an occluded environment of altered chemistry in which the steel matrix is unstable^{19,21}; (5) stable pitting ensues. A schematic representation of these processes is shown in Fig. 2.

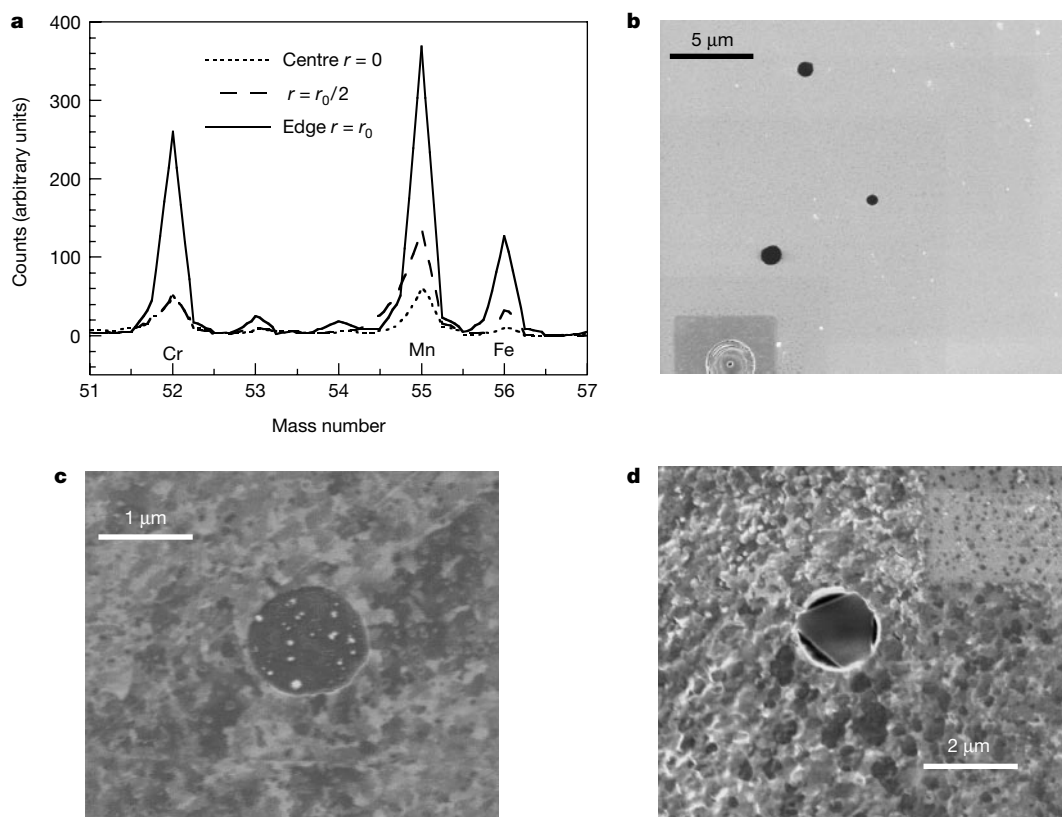


Figure 3 Variation in chemistry within MnS particles in 316F stainless steel. **a**, SIMS spectra obtained at sites inside a sulphide inclusion (radius, r_0). Substantial variations in composition are observed indicating the possible formation of other phases within the inclusion. Beam width, 100 nm; current, 100 pA; energy, 30 keV. **b**, Secondary electron image indicating the distribution of inclusions (small black circles) in the sample. The inclusion at the bottom left of the image was previously analysed. As can be seen, even in a small region there are many possible corrosion initiation sites. Current, 10 pA; energy,

30 keV; scale bar, 5 μm . **c**, Ionic image of an inclusion at mass number 52 (Cr). Areas rich in Cr are clearly evident (bright circles within the inclusion and outer ring) in agreement with observations shown in **a**. Current, 10 pA; energy, 30 keV; scale bar, 1 μm . **d**, Secondary electron image indicating clear crystallographic features and voids within the inclusion itself. Several inclusions showed this type of structure. Current, 11 pA; energy, 30 keV; scale bar, 2 μm .

To support the suggestion that local chemical changes are induced at the inclusion, we find significant compositional variations within the inclusions themselves as can be seen in the mass spectra (Fig. 3a). Several inclusions show an outer ring that is rich in Mn and Cr. In addition, we find that many inclusions show clear crystallographic features, localized high concentration of Cr and evidence of voids within the inclusion (Fig. 3c, d).

We intend to quantify the depletion effects and develop methods of assessing such local changes in and around second-phase precipitates in this and other heterogeneous alloy systems.

The data we have shown provides direct evidence for the mechanism that operates in the pitting corrosion of stainless steels that is consistent with the large body of corrosion literature²². A localized zone exists that is known to display enhanced electrochemical activity and we relate this directly to local chemical variations. This critical zone is a region that is on the order of only 200–400 nm across, yet is responsible for some of the most catastrophic failures of metallic structures. Our findings also focus on the steel processing conditions that allow the set-up of local depleted zones, and hence suggest methods of minimizing the likelihood of pitting corrosion. □

Received 5 November; accepted 19 December 2001.

- Williams, D. E., Westcott, C. & Fleischmann, M. Stochastic models of pitting corrosion of stainless steels. 1. Modeling of the initiation and growth of pits at constant potential. *J. Electrochem. Soc.* **132**, 1796–1804 (1985).
- Eklund, G. S. Initiation of pitting at sulfide inclusions in stainless steel. *J. Electrochem. Soc.* **121**, 467–473 (1974).
- Wranglen, G. Pitting and sulphide inclusions in steel. *Corrosion Sci.* **14**, 331–349 (1974).
- Williams, D. E. & Zhu, Y. Y. Explanation for initiation of pitting corrosion of stainless steels at sulfide inclusions. *J. Electrochem. Soc.* **147**, 1763–1766 (2000).
- Galvele, J. R. Transport processes and the mechanism of pitting of metals. *J. Electrochem. Soc.* **123**, 464–474 (1976).
- Hoar, T. P., Mears, D. C. & Rothwell, G. P. The relationships between anodic passivity, brightening and pitting. *Corrosion Sci.* **5**, 279–289 (1981).
- Chao, C. Y., Lin, L. F. & Macdonald, D. D. A point defect model for anodic passive films. *J. Electrochem. Soc.* **128**, 1187–1194 (1981).
- Uhlig, H. H. Adsorbed and reaction-production films on metals. *J. Electrochem. Soc.* **97**, 215C (1950).
- Sato, N. A. Theory for breakdown of anodic oxide films on metals. *Electrochim. Acta* **16**, 1683–1692 (1971).
- Richardson, J. A. & Wood, G. C. Study of the pitting corrosion of Al by scanning electron microscopy. *Corrosion Sci.* **10**, 313–323 (1970).
- Suter, T. & Bohni, H. Microelectrodes for corrosion studies in microsystems. *Electrochim. Acta* **47**, 191–199 (2001).
- Lott, S. E. & Alkire, R. C. The role of inclusions on initiation of crevice corrosion of stainless steel. 1. Experimental studies. *J. Electrochem. Soc.* **136**, 973–979 (1989).
- Marcus, P., Teissier, A. & Oudar, J. The influence of sulphur on the dissolution and passivation of a NiFe Alloy. 1. Electrochemical and radio tracer measurements. *Corrosion Sci.* **24**, 259–268 (1984).
- Baker, M. A. & Castle, J. E. The initiation of pitting corrosion at MnS inclusions. *Corrosion Sci.* **34**, 667–682 (1993).
- Ryan, M. P., Laycock, N. J., Newman, R. C. & Isaacs, H. S. The pitting behaviour of thin film FeCr alloys in hydrochloric acid. *J. Electrochem. Soc.* **145**, 1566–1571 (1998).
- Stewart, J. & Williams, D. E. The initiation of pitting corrosion on austenitic stainless steels: on the role and importance of sulphide inclusions. *Corrosion Sci.* **33**, 457–474 (1992).
- Monnartz, P. Iron–chromium alloys with special consideration of resistance to acids. *Metallurgie* **8**, 161–176 (1911).
- Williams, D. E., Newman, R. C., Song, Q. & Kelly, R. G. Passivity breakdown and pitting corrosion of binary alloys. *Nature* **350**, 216–219 (1991).
- Webb, E. G., Suter, T. & Alkire, R. C. Microelectrochemical measurements of the dissolution of single MnS inclusions, and the prediction of critical conditions for pit initiation on stainless steel. *J. Electrochem. Soc.* **148**, B186–B195 (2001).
- Williams, D. E., Mohiuddin, T. F. & Zhu, Y. Elucidation of a trigger mechanism for pitting corrosion of stainless steels using sub-micron resolution SECM and photoelectrochemical microscopy. *J. Electrochem. Soc.* **145**, 2664–2672 (1998).
- Brossia, C. S. & Kelly, R. G. Influence of sulfur content and bulk electrolyte composition on crevice corrosion initiation of austenitic stainless steel. *Corrosion* **54**, 145–154 (1998).
- Frankel, G. S. Pitting corrosion of metals: a review of the critical factors. *J. Electrochem. Soc.* **145**, 2186–2197 (1998).

Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council, UK.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.P.R. (e-mail: m.p.ryan@ic.ac.uk).

Poleward heat transport by the atmospheric heat engine

Leon Barry, George C. Craig & John Thuburn

Department of Meteorology, University of Reading, Earley Gate, Whiteknights, Reading RG6 6BB, UK

The atmospheric heat transport on Earth from the Equator to the poles is largely carried out by the mid-latitude storms. However, there is no satisfactory theory to describe this fundamental feature of the Earth's climate^{1,2}. Previous studies have characterized the poleward heat transport as a diffusion by eddies of specified horizontal length and velocity scales, but there is little agreement as to what those scales should be^{3–7}. Here we propose instead to regard the baroclinic zone—the zone of strong temperature gradients and active eddies—as a heat engine which generates eddy kinetic energy by transporting heat from a warmer to a colder region. This view leads to a new velocity scale, which we have tested along with previously proposed length and velocity scales, using numerical climate simulations in which the eddy properties have been varied by changing forcing and boundary conditions. The experiments show that the eddy velocity varies in accordance with the new scale, while the size of the eddies varies with the well-known Rhines β -scale. Our results not only give new insight into atmospheric eddy heat transport, but also allow simple estimates of the intensities of mid-latitude storms, which have hitherto only been possible with expensive general circulation models.

The poleward eddy heat transport in a planetary atmosphere can be written as the vertical integral of $\overline{v'T'}$ = $k|\overline{v'}||T'|$, where overbar means a zonal average, prime means a departure from the zonal average, $||$ means a root-mean-square magnitude, and k is the correlation between the meridional eddy velocity v' and the eddy temperature perturbation T' . If it is assumed that $|T'|$ = $-L_{\text{disp}}\overline{T}_y$, where $\overline{T}_y$ is the poleward temperature gradient and L_{disp} is a typical meridional fluid parcel displacement, then $\overline{v'T'}$ = $-D\overline{T}_y$, where the transport coefficient D = $L_{\text{disp}}|\overline{v'}|$. These flux-gradient theories generally assume that the correlation k is constant^{3–5}, and attempt to predict characteristic scales L_{disp} and $|\overline{v'}|$ that will occur on a given zonal-mean state. We found the constant- k assumption to hold to a good approximation for the experiments discussed below; k was in the range 0.2–0.35 despite variations in the heat flux of about two orders of magnitude⁸. A constant value k = 0.25 is used in all subsequent calculations. It is currently an open question whether a flux-gradient relation can give useful quantitative predictions of eddy heat transport, because it requires the assumption that L_{disp} is much smaller than the width of the baroclinic zone⁹. The results below imply that this is true for the Earth.

Table 1 Summary of diffusivity parameters

Theory	L	$ \overline{v'} $	D
Stone ⁴	L_D	$\overline{u}$	$k\overline{u}L_D$
Green ³	L_{zone}	$\overline{u}L_{\text{zone}}/L_D$	$k\overline{u}L_{\text{zone}}^2/L_D$
Haine and Marshall ⁷	L_{zone}	$\overline{u}$	$k\overline{u}L_{\text{zone}}$
Held and Larichev ⁶	L_β	$\overline{u}\gamma^{-1}$	$k\overline{u}L_D\gamma^{-2}$
Branscome ^{5,13}	L_D	$\overline{u}$	$k\overline{u}L_D$
	$(0.48 + 1.48\gamma)$	$(0.48 + 1.48\gamma)$	$(0.48 + 1.48\gamma)$
This work	L_β	$\left(\frac{eaT_yq}{T_0}\right)^{2/5} \left(\frac{2}{\beta}\right)^{1/5}$	$\left(\frac{eaT_yq}{T_0}\right)^{3/5} \left(\frac{2}{\beta}\right)^{4/5}$

$L_D = NH/f$, $L_\beta = (2|\overline{v'}|/\beta)^{1/2}$ and $\gamma = \beta_0 H/(f\theta_0)$.

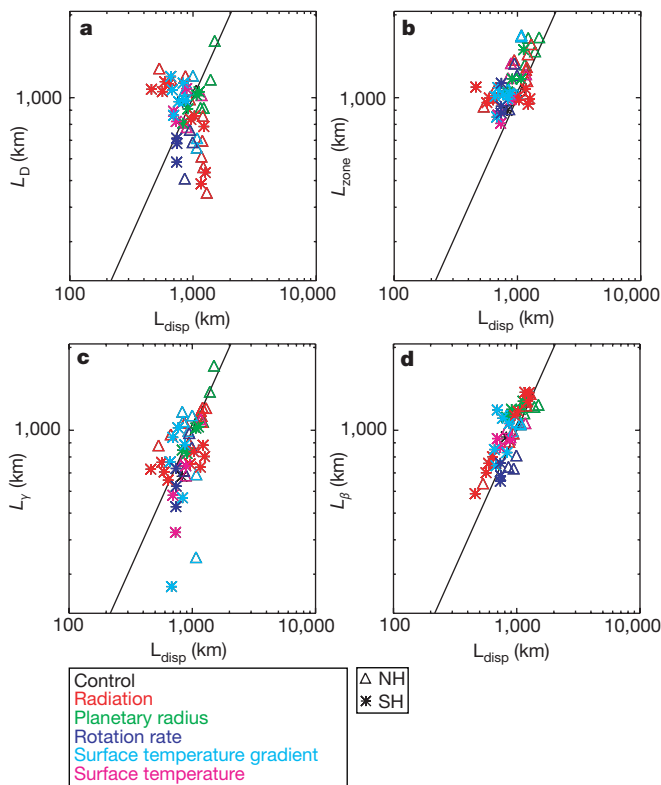


Figure 1 Predicted length scales versus diagnosed eddy length scale $L_{\text{disp}} = \overline{v' T'} / (k \overline{T'} |v'|)$. The four predicted length scales are: **a**, $L_D = NH/f$, the radius of deformation; **b**, L_{zone} , the size of the baroclinic zone; **c**, $L_\gamma = L_D / (0.48 + 1.48\gamma)$ where $\gamma = (\beta \theta_z H_s) / (f \theta_\phi)$; and **d**, $L_\beta = (2|v'|/\beta)^{1/2}$ is the Rhines β -scale. The four length scales have been normalized so that they are equal to L_{disp} for the Northern Hemisphere of the control climate. Symbol type represents the hemisphere, colour represents the type of experiment. NH, Northern Hemisphere; SH, Southern Hemisphere.

Different assumptions about the dominant physical processes in the atmosphere lead to very different predictions for L_{disp} and $|v'|$. The turbulent eddies that transport heat extract energy from the mean flow primarily through baroclinic instability, suggesting that a plausible prediction for L_{disp} may be obtained from the size of the fastest-growing wave found in a linear model of this process. The fastest-growing wave in the Eady model¹⁰, one of the simplest analytic models of baroclinic instability, has zonal length scale proportional to the radius of deformation $L_D = NH/f$, where N is the buoyancy frequency, H is the depth of the domain (in this case the troposphere) and f is the Coriolis parameter, defined as $2\Omega \sin \phi$, where Ω is the Earth's rotation rate, and ϕ is the latitude. If nonlinear interactions are assumed to make the eddies circular, then the meridional scale will be proportional to L_D too⁴. An alternative model of baroclinic instability is the Charney model¹¹, in which an approximation to the fastest-growing mode has a horizontal length scale^{5,12,13} $L_\gamma = (NH_s/f) / (0.48 + 1.48\gamma)$, where $\gamma = \beta \theta_z H_s / f \theta_\phi$. In this expression, H_s is the density scale height, $\beta = f_y$ is the northward gradient of f , and θ_z and θ_ϕ are the vertical and horizontal derivatives of the potential temperature $\theta = T(p_0/p)^\kappa$, where p_0 is a reference pressure of 1,000 mbar and $\kappa = R/c_p$ is the ratio of the gas constant for dry air to the specific heat at constant pressure. However, the main energy-containing eddy scale will not necessarily be that of the fastest-growing baroclinic mode. Other candidate length scales are suggested by geostrophic turbulence theory, which postulates a cascade of energy to larger horizontal scales^{14,15} that may dominate the heat transport. The largest scale to which a cascade could extend is L_{zone} , the size of

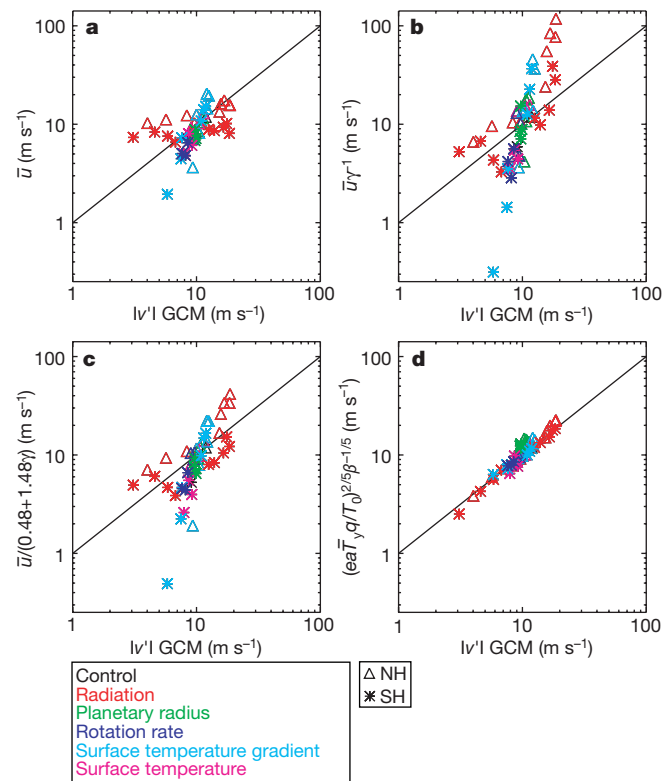


Figure 2 Predicted velocity scales versus diagnosed eddy velocity scale $|v'|$. The four predicted velocity scales are $\bar{u}$ (**a**), $\bar{u}\gamma^{-1}$ (**b**), $\bar{u}/(0.48 + 1.48\gamma)$ (**c**) and $(e\bar{a}T_0/q/T_0)^{2/5} (2/\beta)^{1/5}$ (**d**) (Symbols are defined in the text). The four velocity scales have been normalized so that they are equal to $|v'|$ for the Northern Hemisphere of the control climate. Symbol type and colour have same meaning as in Fig. 1. GCM, general circulation model.

the baroclinic zone^{3,16}. A final possibility is that the cascade to larger scales may be halted before L_{zone} at the Rhines β -scale $L_\beta = (2|v'|/\beta)^{1/2}$, the scale at which the Rossby wave restoring mechanism becomes comparable to advective nonlinearity¹⁷.

A plausible estimate for the eddy velocity scale is that it is proportional to the speed of the zonal mean flow^{7,18}, that is, $|v'| \propto \bar{u}$. However, it has also been suggested that there should be an equipartition between eddy potential energy and eddy kinetic energy^{4,6}, which can be shown to lead to the relation $|v'| \propto \bar{u}(L_{\text{disp}}/L_D)$. These scales depend upon the local properties of the zonal mean state.

We propose an alternative velocity scale that comes from a thermodynamic constraint on the baroclinic zone as a whole. This is obtained by considering the baroclinic zone (rather than the global atmosphere, as previously proposed¹⁹) as a heat engine that obtains its kinetic energy in the process of transporting heat from a warmer to a colder region. The rate of generation and dissipation of eddy kinetic energy is given by $\epsilon = e(\delta T/T_0)q$, where q is the rate of energy transport out of the tropics, calculated as the net diabatic (radiative plus surface flux) cooling per unit mass averaged over the baroclinic zone, $\delta T/T_0$ is the maximum possible thermodynamic efficiency, δT is the temperature difference between the regions where the heat is put in and where it is extracted, and e is the utilization coefficient (assumed constant), which measures the fraction of the generated kinetic energy used by heat-transporting eddies. If the heat-transporting eddies exist in an inertial range, for example in the postulated upscale energy cascade, then the average properties of the flow will depend only on the dissipation rate

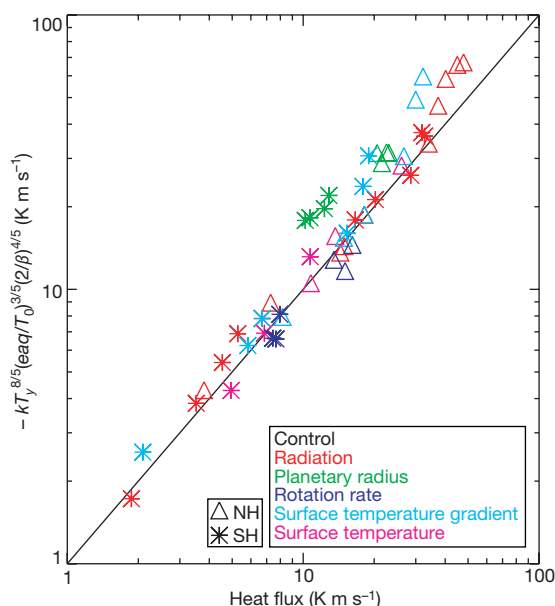


Figure 3 Parametrized heat flux versus heat flux diagnosed directly from the GCM. The heat flux predicted by the new parametrization is $\bar{v}'T'_{\text{param}} = kT_y^{8/5}(eaq/T_0)^{3/5}(2/\beta)^{4/5}$. The predicted flux has been normalized so that it is equal to the GCM flux for the Northern

Hemisphere of the control climate. The symbol type and colour have the same meanings as in Fig. 1.

and the wavelength. This implies a scaling for the velocity of $|v'| \propto (\epsilon L_{\text{disp}})^{1/3}$. If we compare the two expressions involving ϵ and assume that $\delta T \propto a|\bar{T}_y|$ where a is the radius of the Earth, and anticipate the result that $L_{\text{disp}} \propto L_\beta$, then this leads to the expression $|v'| \propto (ea|\bar{T}_y|q/T_0)^{2/5}(2/\beta)^{1/5}$.

The various velocity and length scales introduced above have been used to define a number of heat flux parametrizations, as summarized in Table 1. The validity of these parametrizations has been explored in many studies using observations of the Earth's climate²⁰, numerical models of idealized atmospheres^{16,21–24} and oceans^{7,25}. But it has proved difficult to obtain a conclusive test, owing to the need to explore a very wide range of parameters, in a system with sufficient physical realism. Here we show results of tests using a general circulation model (GCM) of the Earth's atmosphere²⁶. The model includes orography and a full set of physical parametrizations, and is forced by fixed sea surface temperatures and perpetual-January radiative forcing. The candidate length and velocity scales were calculated for model climates with different forcings and boundary conditions (see Methods).

Figure 1 is a plot of the various candidate length scales against a measure of the mean fluid parcel displacement, $L_{\text{disp}} = -\bar{v}'T'/(k|v'|T_y)$. L_β is the best correlated with L_{disp} (correlation 0.79), although some skill is evident for L_{zone} with a correlation of 0.63. Figure 2 is a plot of the candidate velocity scales against the magnitude of eddy velocity measured in the GCM. Clearly our velocity parametrization is the most accurate with a correlation of 0.96. (The values of correlations quoted here are for the log-log correlation, which are more appropriate for the large range of scales involved. The statistical significance of the correlations was estimated as >99.9% with a Student *t*-test.)

Combining the length scale L_β with our velocity expression leads to a new parametrization for the heat flux $\bar{v}'T' = kT_y^{8/5}(eaq/T_0)^{3/5}(2/\beta)^{4/5}$. Figure 3 is a plot of this parametrization against the horizontal eddy heat flux measured in the GCM runs. The correlation between them is 0.97. A physical picture that is consistent with this scaling is of an upscale turbulent energy cascade that is halted at L_β where nonlinear scale interactions are suppressed.

Such a cascade has been proposed by Held and Larichev⁶, although they suggested that the throughput of kinetic energy was constrained by baroclinic production, rather than by the heat engine efficiency. Preliminary results²⁷ suggest that the transport coefficient implied by the above expression for horizontal heat flux may also serve as the basis for expressions for vertical heat flux and for transport of moisture.

As noted earlier, an important prerequisite for the application of a flux-gradient parametrization is that the baroclinic zone be wider than the displacement length⁹, L_{disp} . Because the baroclinic zone width in the present climate is about four times L_β , it is reasonable to expect a flux-gradient relationship to hold when averaged over this zone (Fig. 3). Calculations with averaging over latitude bands comparable in size to L_β show similar collapse of the data to that seen in Fig. 3, but there is a suggestion of a systematic variation of slope with latitude, whose significance remains unclear.

A parametrization for eddy heat flux is useful not only for the insight it provides into the workings of the atmosphere, but also for practical climate modelling at a level intermediate between simple energy-balance models and complex GCMs. The heat engine description proposed here gives a thermodynamic constraint on the eddies that allows prediction of the heat transport and pole to Equator temperature difference as well as an estimate of the intensity of mid-latitude storms. Such predictions are not possible with energy-balance models that do not attempt to describe eddy motions at all, and are very expensive with GCMs that explicitly describe every individual eddy. □

Methods

Numerical experiments were performed on the Reading intermediate GCM²⁶ run at T42 spectral horizontal resolution with 33 vertical levels. This resolution is considered sufficient to represent the global transports by eddies²⁸. Each experiment was for 120 d with diagnostics taken from the last 90 d. The experiments were as follows. Radiation experiments: radiative heating rates multiplied everywhere by a constant factor of 0.01, 0.1, 0.5, 1, 2, 3, 5, 10 or 20. Planetary radius experiments: radius increased by a factor of 1.2, 1.4, 1.5, 1.6 or 1.7. Rotation rate experiments: planetary rotation rate increased by a factor of 1.4, 1.6 or 1.7. Temperature gradient experiments: surface temperature gradient multiplied by 0.3, 0.6, 1.2, 1.4, 1.8 or 2. Temperature experiments: surface temperature

changed uniformly by +5 K, −10 K or −20 K. Additional experiments, not presented here, demonstrated insensitivity of the simulated heat fluxes to changes in surface fluxes of heat, moisture and momentum, or in scale-selective dissipation.

Except as noted below, all quantities used to create Figs 1–3 are zonal means, averaged over the mid-latitude region defined as 1,000–200 mbar, 28°–68° latitude, for each hemisphere. The results were found to be insensitive to variations in the definition of the mid-latitude region. In particular, this height is sufficient to include essentially the whole troposphere and all the eddy activity. It should be noted that the average meridional temperature gradient is closely constrained by the fixed surface temperatures, and could be replaced by the surface temperature gradient in the calculations with only a modest loss in accuracy. The diabatic forcing term q includes both the radiative heating averaged over the mid-latitude region and the surface fluxes of sensible heat, although the contribution of surface fluxes turns out to be small. The height scales H and H_0 , used to calculate the radius of deformation L_D , and the Charney length scale L_{γ} , were taken to be constant at 8 km as they did not vary substantially in any of the experiments. In calculating L_{γ} , the vertical averages were weighted¹³ by a factor $e^{-z/D}$, where $D = H_0/(0.48 + 1.48\gamma)$. The latitude of the zonal wind maximum, which did not vary substantially between experiments, was used to calculate f and β . The size of the baroclinic zone, L_{zone} , was defined as the distance between the two latitudes where the strength of the vertically averaged zonal-mean zonal wind first drops to half of its maximum value. In all calculations the correlation coefficient k was given a constant value of 0.25, while the utilization coefficient e was given a constant value of 0.75.

Received 1 November; accepted 18 December 2001.

- Pierrehumbert, R. T. & Swanson, K. L. Baroclinic instability. *Annu. Rev. Fluid Mech.* **27**, 419–467 (1995).
- Held, I. M. The macroturbulence of the troposphere. *Tellus A* **51**, 59–70 (1999).
- Green, J. S. Transfer properties of the large scale eddies and the general circulation of the atmosphere. *Q. J. R. Meteorol. Soc.* **96**, 157–185 (1970).
- Stone, P. H. A simplified radiative-dynamical model for the static stability of rotating atmospheres. *J. Atmos. Sci.* **29**, 405–418 (1972).
- Branscome, L. E. A parameterization of transient eddy heat flux on a beta plane. *J. Atmos. Sci.* **41**, 2508–2521 (1983).
- Held, I. M. & Larichev, V. D. A scaling theory for horizontally homogeneous, baroclinically unstable flow on a beta plane. *J. Atmos. Sci.* **53**, 946–952 (1996).
- Haine, T. W. N. & Marshall, J. Gravitational, symmetric and baroclinic instability of the ocean mixed layer. *J. Phys. Oceanogr.* **28**, 634–658 (1998).
- Barry, L., Thuburn, J. & Craig, G. C. GCM tests of some possible dynamical constraints on the mid-latitude atmosphere: The v' – T' correlation, PV homogenisation and the dividing isentrope. *Q. J. R. Meteorol. Soc.* (submitted).
- Pavan, V. & Held, I. M. The diffusive approximation for eddy fluxes in baroclinically unstable jets. *J. Atmos. Sci.* **53**, 1262–1272 (1996).
- Eady, E. T. Long waves and cyclone waves. *Tellus* **1**, 33–52 (1949).
- Charney, J. G. The dynamics of long waves in a baroclinic westerly current. *J. Meteorol.* **4**, 135–162 (1947).
- Held, I. M. The vertical scale of an unstable baroclinic wave and its importance for eddy heat flux parameterisation. *J. Atmos. Sci.* **35**, 572–576 (1978).
- Stone, P. H. & Yao, M.-S. Development of a 2-dimensional zonally averaged statistical-dynamic model. 3. The parameterisation of eddy fluxes of heat and moisture. *J. Clim.* **3**, 726–740 (1990).
- Charney, J. G. Geostrophic turbulence. *J. Atmos. Sci.* **28**, 1087–1095 (1971).
- Salmon, R. *Lectures on Geophysical Fluid Dynamics* (Oxford Univ. Press, Oxford, 1998).
- Larichev, V. D. & Held, I. M. Eddy amplitudes and fluxes in a homogeneous model of fully developed baroclinic instability. *J. Phys. Oceanogr.* **25**, 2285–2297 (1995).
- Rhines, P. B. Waves and turbulence on a beta-plane. *J. Fluid. Mech.* **69**, 417–443 (1975).
- James, I. N. *Introduction to Circulating Atmospheres* (Cambridge Univ. Press, Cambridge, 1994).
- Golitsyn, G. S. A similarity approach to the general circulation of planetary atmospheres. *Icarus* **13**, 1–24 (1970).
- Stone, P. H. & Miller, D. A. Empirical relations between seasonal changes in meridional temperature gradients and meridional fluxes of heat. *J. Atmos. Sci.* **37**, 1708–1721 (1980).
- Vallis, G. K. Numerical studies of eddy transport properties in eddy-resolving and parameterised models. *Q. J. R. Meteorol. Soc.* **114**, 183–204 (1988).
- Stone, P. H. & Branscome, L. Diabatically forced, nearly inviscid eddy regimes. *J. Atmos. Sci.* **49**, 355–367 (1992).
- Panetta, R. L. Zonal jets in wide baroclinically unstable regions: persistence and scale selection. *J. Atmos. Sci.* **29**, 2073–2106 (1993).
- James, I. N. Two parameterisations of the temperature flux due to baroclinic waves. *Q. J. R. Meteorol. Soc.* **123**, 1–16 (1997).
- Visbeck, M., Marshall, J. & Haine, T. Specification of eddy transfer coefficients in coarse resolution ocean circulation models. *J. Phys. Oceanogr.* **27**, 381–402 (1997).
- Forster, P. M. de F., Blackburn, M., Glover, R. & Shine, K. P. An examination of climate sensitivity for idealised climate change experiments in an intermediate general circulation model. *Clim. Dyn.* **16**, 833–849 (2000).
- Barry, L. *Predicting Eddy Heat Transport in the Troposphere* Thesis, Univ. Reading (2000).
- Boer, G. & Denis, B. Numerical convergence of the dynamics of a GCM. *Clim. Dyn.* **13**, 359–374 (1997).

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.C.C. (e-mail: G.C.Craig@Reading.ac.uk).

Mid-mantle deformation inferred from seismic anisotropy

James Wookey*, J.-Michael Kendall* & Guilhem Barruol†

* School of Earth Sciences, University of Leeds, Woodhouse Lane, Leeds LS2 9JT, UK

† CNRS, Université Montpellier, 34095 Montpellier, Cedex 05, France

With time, convective processes in the Earth's mantle will tend to align crystals, grains and inclusions. This mantle fabric is detectable seismologically, as it produces an anisotropy in material properties—in particular, a directional dependence in seismic-wave velocity. This alignment is enhanced at the boundaries of the mantle where there are rapid changes in the direction and magnitude of mantle flow¹, and therefore most observations of anisotropy are confined to the uppermost mantle or lithosphere^{2,3} and the lowermost-mantle analogue of the lithosphere, the D' region⁴. Here we present evidence from shear-wave splitting measurements for mid-mantle anisotropy in the vicinity of the 660-km discontinuity, the boundary between the upper and lower mantle. Deep-focus earthquakes in the Tonga–Kermadec and New Hebrides subduction zones recorded at Australian seismograph stations record some of the largest values of shear-wave splitting hitherto reported. The results suggest that, at least locally, there may exist a mid-mantle boundary layer, which could indicate the impediment of flow between the upper and lower mantle in this region.

Seismic anisotropy in the upper 200 km of the Earth's mantle is primarily attributed to the preferred alignment of olivine crystals which have deformed by dislocation creep⁵. The origin of anisotropy at greater depths is more speculative, but there is evidence for anisotropy in the transition zone in some regions^{6–8}, but not in others^{9,10}. In an effort to reconcile discrepancies in global velocity models derived from body-wave travel times and normal-mode observations, Montagner and Kennett¹¹ allowed both anisotropy and attenuation in a joint inversion of these data sets. Their final model shows significant levels of anisotropy in the uppermost and lowermost mantle, but also in the vicinity of the 660-km discontinuity (hereafter referred to as the '660'). This motivated an investigation of mid-mantle anisotropy on a regional scale. Here we investigate shear-wave splitting in deep-focus events that image a region below the Australian plate (Fig. 1).

Stations in Australia are ideal for investigating near-source anisotropy, as studies have shown that they exhibit very little, if any, receiver-side shear-wave splitting^{12–14} (see Supplementary Information for a summary of observations). For example, 52 SKS measurements with good azimuthal coverage at the station CAN (see Fig. 1 for location) show that shear waves that are travelling nearly vertically are not split while crossing the Australian lithosphere beneath this station¹². In contrast, we find that deep-focus events from the Tonga–Kermadec and New Hebrides subduction zones show very large degrees of shear-wave splitting at this and four other Australian stations (Fig. 2), suggesting anisotropy deeper in the mantle, away from the receiver.

We made splitting measurements from 92 events, at epicentral distances of 24° to 59° from the Australian stations, using the method of ref. 15, which estimates the time separation between the fast and slow shear wave, δt , and the polarization of the fast shear wave at the receiver, ϕ . This method attempts to remove the anisotropy-induced splitting by minimizing the shear-wave signal in the direction perpendicular to the polarization direction of the shear wave before entering the anisotropic region¹⁶. A grid search over δt and ϕ is used to estimate the splitting parameters, and a statistical F -test is used to assess errors. The correction for splitting

should produce a linear S-wave particle motion, thus providing a further measure of confidence in the results. Of 164 splitting measurements, 66 gave very convincing results—that is, the error in δt is less than 0.5 s, and the error in ϕ is less than 10° . In an effort to isolate mid-mantle anisotropy, we further restricted our study to the 30 events deeper than 300 km which gave 35 high-quality splitting estimates (see Supplementary Information; 75% of the events are greater than 500 km deep). The magnitude of splitting for these events ranges from 0.6 s to 7.1 s. Many measurements show splitting in excess of 4 s (Fig. 2), and suggest either very high degrees of localized anisotropy or wave propagation through a more moderately anisotropic region of large extent. It should be noted that the maximum free-surface incidence angle for our data set is less than 32° , thus avoiding the effects of waveform distortion due to free- and near-surface coupling.

The azimuthal ray coverage at the midpoint between source and receiver spans a 60° region centred around 260° . The polarization of the fast shear wave is roughly aligned with the transverse component, but there is some scatter in this (back-azimuth $-\phi = 108^\circ \pm 31^\circ$). Although the azimuthal ray coverage is not complete, the results suggest a transversely isotropic symmetry with the symmetry axis in the vertical plane, perpendicular to the ray direction. For horizontally travelling rays, this would imply a horizontally polarized fast shear wave (that is, SH leads SV).

To help guide interpretations of these observations, we model wave propagation through an anisotropic slab region using ray theory¹⁷. The linear slab extends to a depth of 660 km and has a 60° dip angle. Anisotropy in the deeper parts of the slab may be due to the alignment of metastable olivine¹⁸, or the preferred alignment of akimotoite, a polymorph of enstatite, which may exist under slab pressures and temperatures¹⁹. Alternatively, transition-zone deformation above the slab may align its dominant minerals, wadsleyite and ringwoodite. Finally, anisotropy below the '660' may be due to the alignment of lower-mantle minerals such as perovskite, periclase and/or stishovite, all of which are highly anisotropic²⁰. Perovskite is the most likely candidate as it constitutes nearly 80% of the minerals in this region, but an alignment mechanism for perovskite is still uncertain^{21,22}. Both experimental measurements²³ and first-principles calculations^{24,25} suggest that perovskite is only mildly orthor-

hombic in symmetry and can be well approximated as being transversely isotropic. Unfortunately, these elastic constants predict that for horizontally travelling waves, vertically polarized shear waves are faster than horizontally polarized shear waves. Alternatively, the anisotropy may not be due to crystal alignment, but rather to the horizontal alignment of tubular inclusions, as has been suggested for the lowermost mantle⁴.

Figure 2 compares predicted splitting for a variety of models and the observed splitting values. Consistent with the observations, the anisotropy is constrained to have a fast horizontally polarized shear wave. Very high degrees of anisotropy, distributed throughout the slab, are required to explain the observations with slab anisotropy. It is virtually impossible to explain 6 seconds of splitting in the deepest events (>600 km) with slab anisotropy. An absence of slab anisotropy is further suggested from an analysis of depth-dependent splitting for vertically travelling shear waves beneath Tonga, which shows no evidence of azimuthal anisotropy below 400 km (ref. 10). A lack of SKS splitting does not necessarily mean that the uppermost mantle beneath the receivers is isotropic. Transverse isotropy anywhere in the mantle will not split vertically polarized SKS phases, but will split an arbitrarily polarized S wave. However, Fig. 2 shows that the splitting for an uppermost mantle with 4% anisotropy²⁶ cannot explain the results. Similarly, they cannot be explained with anisotropy confined to the transition zone, a conclusion reinforced by the fact that ringwoodite, the dominant mineral between depths of 520 and 660 km, is thought to be only very mildly anisotropic²⁰. It is difficult to explain the splitting with combinations of transverse isotropy in the uppermost mantle and in transition-zone regions. A problem with models where anisotropy is confined to the upper mantle is that they predict very large amounts of splitting at near offsets and little at large offsets, an effect not seen in the data. Such anisotropy may contribute to the splitting, but cannot explain the observations.

The modelling shows that moderate amounts of anisotropy in the lower mantle generate large amounts of splitting owing to long horizontal ray-paths below the '660' at these epicentral distances. Assuming that the anisotropy is confined to a layer 100 km below the '660', the average anisotropy magnitude is 3.0%. However, Fig. 1 shows that there is spatial variation in this estimate, with the largest

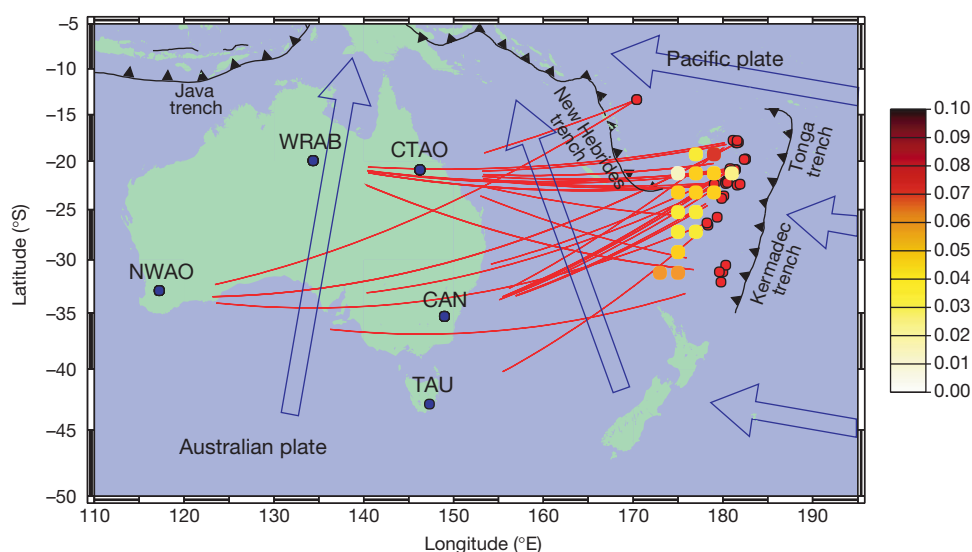


Figure 1 Event-station combinations used to study mid-mantle anisotropy. Small red dots show events; blue dots show stations; red swaths show corresponding lower-mantle ray-paths. Subduction zones and directions of absolute plate motion are also indicated. The larger circles near the events are colour-coded averages (minimum 3 hits) over a 2° radius surface region of the anisotropy magnitude required to explain observed shear-

wave splitting, plotted at the point a ray enters the lower mantle. This is calculated assuming that the anisotropy is distributed throughout a 100-km-thick layer just below the 660-km discontinuity (the '660'). Colour scale bar to the right indicates the magnitude of the anisotropy (fractional difference between slow and fast shear wave velocity).

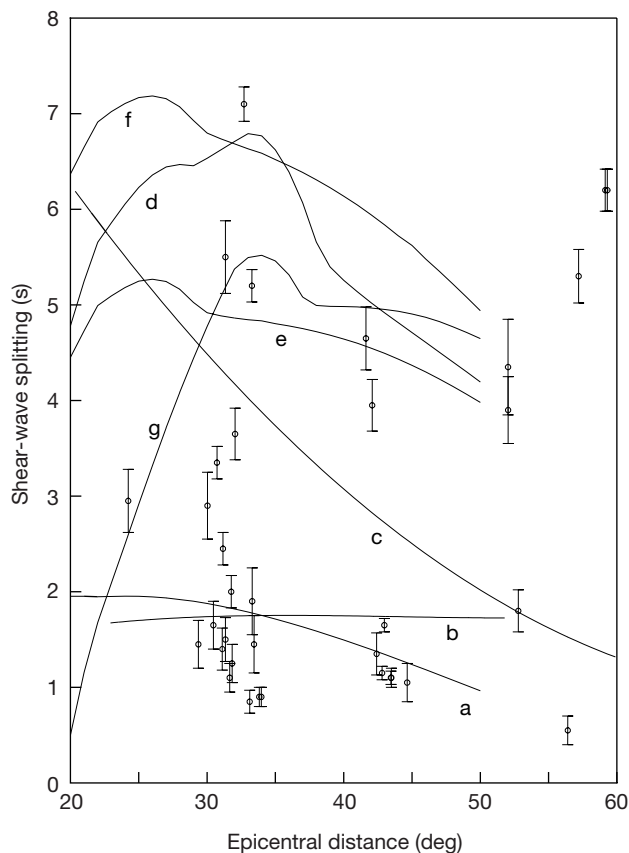


Figure 2 Shear-wave splitting versus epicentral distance. Circles with error bars show observations and estimated uncertainty. Solid lines show predictions for a 500-km-deep source in models with a fast horizontally polarized shear wave: trace a, 4% transverse-isotropy in the uppermost 210 km of upper mantle; b, 5% anisotropy in a subducted slab that extends to a depth of 660 km; c, 5% anisotropy confined to a 100-km-thick layer immediately beneath the '660'; d, anisotropy that grades from 3% to 1% between the '660' and 900 km; e, 2% anisotropy in 100-km-thick layers above and below the '660', and 1% anisotropy in a layer between 760 km and 900 km; f, 4% anisotropy in the uppermost 210 km, 2% anisotropy in 100-km-thick layers above and below the '660', and 1% anisotropy in a layer between 760 km and 900 km; g, anisotropy that grades from 2.5% to 1.5% in a layer between 760 km and 900 km. Although there is some ambiguity as to the best model, only models with anisotropy in the lower mantle can explain the large splitting observations.

magnitude of anisotropy lying in the northernmost and southernmost regions. This north–south variation is also seen in the raw splitting values, with a large range of splitting near 20°S (see Supplementary Information). Alternatively, the magnitude of the anisotropy may grade into the isotropic mantle over a few hundred kilometres, thereby requiring even less anisotropy. Figure 2 also shows that it is difficult accurately to constrain the anisotropy to a layer immediately beneath the '660'. In fact, anisotropy between depths of 760 km and 900 km explains the trend in the large residuals quite well. In summary, the modelling shows that there must be anisotropy below the '660', but not deeper than 900 km, and there may or may not be a contribution from anisotropy above the '660'.

Although there must be anisotropy in the lower mantle, it may still be slab related. Numerical simulations have shown that large deviatoric stresses are transmitted into the lower mantle when a rigid slab encounters an increase in viscosity at the '660' (ref. 27). Large stresses increase the likelihood of dislocation creep mechanisms being active. With time, the associated strains will induce alignment in a broad region below the slab. Perovskite may therefore align with a rotated symmetry axis conformal to the shape of this region. Aligned perovskite rotated more than 30° predicts SH waves faster than SV wave for horizontally travelling S waves.

Another possibility is that the anisotropy is associated with slab material which has broadened and pooled at the '660', before sinking into the deeper mantle. This may be slab material horizontally emplaced on the '660', but our modelling shows that a significant portion of the slab must be well below a depth of 660 km. An alternative is the idea that eclogitic oceanic crust delaminates from the slab, residing in a 'megalith' just below the '660' (ref. 28). This thin crustal layer may thicken appreciably with long-lived subduction into the high-viscosity lower mantle²⁹. Furthermore, it has been argued that basalt may be near its solidus in the uppermost and lowermost parts of the lower mantle³⁰. Thus the anisotropy may be due to the preferred alignment of melt inclusions (a mechanism which generates anisotropy very effectively⁴), which results from shear deformation at the '660'.

Supporting evidence for the anisotropy being confined to a broad region around the base of the slab comes from recent tomographic images for both P and S waves for the Tonga–Kermadec region³¹. This raises the question of to what degree tomographic images obtained assuming isotropy are influenced by anisotropy. Tomographic images also show along-strike variations in the Tonga–

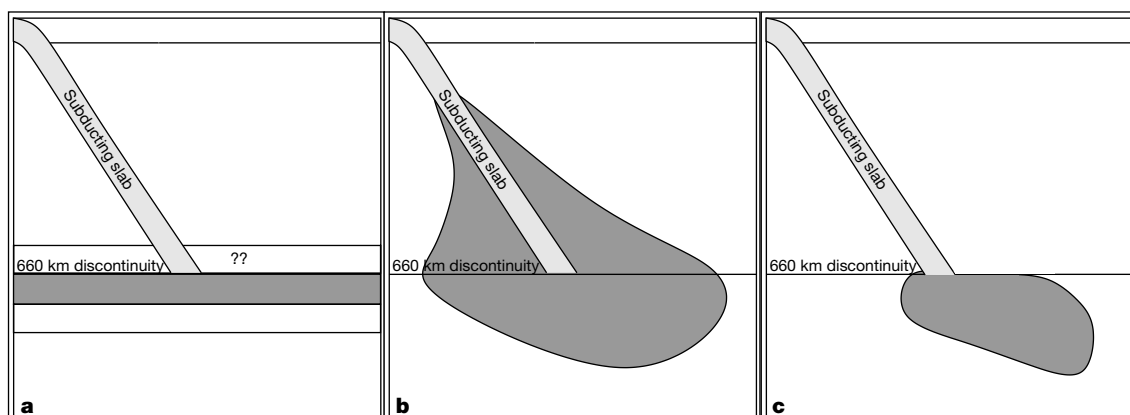


Figure 3 Three models for anisotropy below the 660-km discontinuity. **a**, An anisotropic mid-mantle boundary layer near the '660' that may or may not be a global feature of this boundary. Our results suggest that the magnitude of anisotropy in such a layer must vary

laterally quite significantly. **b**, Slab forces on the surrounding mantle lead to strain-induced anisotropy. **c**, Anisotropy associated with slab material pooling in the lower mantle.

Kermadec slab morphology, with a significant change in slab dip near 25° (ref. 32). We note that it is from this region that we observe the smallest amounts of splitting.

Although the precise origin of the anisotropy is not clear at present, our observations and linked modelling show evidence for anisotropy in the uppermost lower mantle beneath the eastern part of the Australian plate. The anisotropy is probably inhomogeneous, as there appears to be an appreciable north–south variability in its magnitude. There must be large strains in this region, which are probably related to slab interaction with the sharp increase in viscosity at this boundary. Figure 3 summarizes the potential mechanisms that we propose. Our results may help describe to what extent there is an impediment of flow at this boundary between the upper and lower mantle. □

Received 5 September; accepted 31 December 2001.

1. Montagner, J.-P. Where can seismic anisotropy be detected in the Earth's mantle? In boundary layers *Pure Appl. Geophys.* **151**, 223–256 (1998).
2. Silver, P. G. Seismic anisotropy beneath the continents: Probing the depths of geology. *Annu. Rev. Earth Planet Sci.* **24**, 385–432 (1996).
3. Savage, M. K. Seismic anisotropy and mantle deformation: What have we learned from shear wave splitting? *Rev. Geophys.* **37**, 65–106 (1999).
4. Kendall, J.-M. & Silver, P. G. in *The Core-Mantle Boundary Region* (eds Gurnis, M., Wyssession, M., Knittle, E. & Buffett, B.) 97–118 (Geodynamics series 28, American Geophysical Union, Washington DC, 1998).
5. Karato, S. & Wu, P. Rheology of the upper mantle: A synthesis. *Science* **260**, 771–778 (1993).
6. Vinnik, L. P. & Montagner, J.-P. Shear wave splitting in the mantle Ps phase. *Geophys. Res. Lett.* **23**, 2449–2452 (1996).
7. Vinnik, L. P., Chevrot, S. & Montagner, J.-P. Seismic evidence of flow at the base of the upper mantle. *Geophys. Res. Lett.* **25**, 1995–1998 (1998).
8. Tong, C., Gudmundsson, O. & Kennett, B. L. N. Shear wave splitting in refracted waves returned from the upper mantle transition zone beneath northern Australia. *J. Geophys. Res.* **99**, 15783–15797 (1994).
9. Fouch, M. J. & Fischer, K. M. Mantle anisotropy beneath northwest Pacific subduction zone. *J. Geophys. Res.* **101**, 15987–16002 (1996).
10. Fischer, K. & Wiens, D. The depth distribution of mantle anisotropy beneath the Tonga subduction zone. *Earth Planet. Sci. Lett.* **142**, 253–260 (1996).
11. Montagner, J.-P. & Kennett, B. L. N. How to reconcile body-wave and normal-mode reference Earth models. *Geophys. J. Int.* **125**, 229–248 (1996).
12. Barruol, G. & Hoffmann, R. Upper mantle anisotropy beneath Geoscope stations. *J. Geophys. Res.* **104**, 10757–10773 (1999).
13. Clitheroe, G. & Van der Hilst, R. in *Structure and Evolution of the Australian Continent* (ed. Braun, J. et al.) 73–78 (Geodynamics series 26, American Geophysical Union, Washington DC, 1998).
14. Özalaybey, S. & Chen, W. Frequency dependent analysis of SKS/SKKS waveforms observed in Australia: evidence for null birefringence. *Phys. Earth Planet. Inter.* **114**, 197–210 (1999).
15. Silver, P. G. & Chan, W. W. Implications for continental structure and evolution from seismic anisotropy. *Nature* **335**, 34–39 (1988).
16. Wolfe, C. J. & Silver, P. G. Seismic anisotropy of oceanic upper mantle: Shear-wave splitting methodologies and observations. *J. Geophys. Res.* **103**, 749–771 (1998).
17. Kendall, J.-M. & Thomson, C. J. Seismic modelling of subduction zones with inhomogeneity and anisotropy. I: Teleseismic P-wavefront geometry. *Geophys. J. Int.* **112**, 39–66 (1993).
18. Kirby, S. H., Durham, W. B. & Stern, L. A. Mantle phase changes and deep-earthquake faulting in subducting lithosphere. *Science* **252**, 216–225 (1991).
19. Anderson, D. L. Thermally induced phase changes, lateral heterogeneity of the mantle, continental roots and deep slab anomalies. *J. Geophys. Res.* **92**, 13968–13980 (1987).
20. Mainprice, D., Barruol, G. & Ismail, W. B. in *Earth's Deep Interior: Mineral Physics and Tomography from the Atomic to the Global Scale* (eds Karato, S., Forté, A., Masters, T. G. & Stixrude, L.) 237–264 (Geophysical Monographs 117, American Geophysical Union, Washington DC, 2000).
21. Karato, S., Dupas-Bruzek, C. & Rubie, D. Plastic deformation of silicate spinel under the transition-zone conditions of the Earth's mantle. *Nature* **395**, 266–269 (1998).
22. Meade, C. P., Silver, P. G. & Kaneshima, S. Laboratory and seismological observations of lower mantle isotropy. *Geophys. Res. Lett.* **22**, 1293–1296 (1995).
23. Yeganeh-Haeri, A. Synthesis and re-investigation of the elastic properties of single-crystal magnesium silicate perovskite. *Phys. Earth Planet. Inter.* **87**, 111–121 (1994).
24. Oganov, A. R., Brodholt, J. P. & Price, G. D. The elastic constants of MgSiO₃ perovskite at pressures and temperatures of the Earth's mantle. *Nature* **411**, 934–937 (2001).
25. Stixrude, L. in *The Core-Mantle Boundary Region* (eds Gurnis, M., Wyssession, M., Knittle, E. & Buffett, B.) 83–96 (Geodynamics series 28, American Geophysical Union, Washington DC, 1998).
26. Gaherty, J. B. & Jordan, T. H. Lehmann discontinuity as the base of an anisotropic layer beneath continents. *Science* **268**, 1468–1471 (1995).
27. Kuznir, N. J. Subduction body force stresses and viscosity structure at the 410 km and 660 km phase transitions. *Eos* **81**, 1081 (2000).
28. Ringwood, A. E. Role of the transition zone and 660 km discontinuity in mantle dynamics. *Phys. Earth Planet. Inter.* **86**, 5–24 (1994).
29. Mitrovica, J. X. & Forte, A. M. Radial profile of mantle viscosity: Results from the joint inversion of convection and postglacial rebound observations. *J. Geophys. Res.* **102**, 2751–2769 (1997).
30. Hirose, K., Fei, Y., Ma, Y. & Mao, H.-K. The fate of subducted basaltic crust in the Earth's lower mantle. *Nature* **397**, 53–56 (1999).

31. Widiyantoro, S., Kennett, B. L. N. & Van der Hilst, R. Seismic tomography with P and S data reveals lateral variations in the rigidity of deep slabs. *Earth Planet. Sci. Lett.* **173**, 91–100 (1999).
32. Van der Hilst, R. Complex morphology of subducted lithosphere in the mantle beneath the Tonga trench. *Nature* **374**, 154–157 (1995).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com>).

Acknowledgements

We thank K. Fischer, G. Houseman and M. Casey for comments on the manuscript, and K. Fischer for suggesting alternative models to test.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.-M.K. (e-mail: m.kendall@earth.leeds.ac.uk).

A basal troodontid from the Early Cretaceous of China

Xing Xu*, Mark A. Norell†, Xiao-lin Wang*, Peter J. Makovicky‡ & Xiao-chun Wu§

* Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, Beijing 100044, China

† American Museum of Natural History, Central Park West at 79th Street, New York, New York 10024, USA

‡ The Field Museum, 1400 S Lake Shore Drive, Chicago, Illinois 60605, USA

§ Canadian Museum of Nature, Station D, Ottawa, Ontario K1P 6P4, Canada

Troodontid dinosaurs form one of the most avian-like dinosaur groups^{1–5}. Their phylogenetic position is hotly debated, and they have been allied with almost all principal coelurosaurian lineages^{6–13}. Here we report a basal troodontid dinosaur, *Sinovenator changii* gen. et sp. nov., from the lower Yixian Formation of China. This taxon has several features that are not found in more derived troodontids, but that occur in dromaeosaurids and avialans. The discovery of *Sinovenator* and the examination of character distributions along the maniraptoran lineage indicate that principal structural modifications toward avians were acquired in the early stages of maniraptoran evolution.

Theropoda Marsh, 1881
Maniraptora Gauthier, 1986
Troodontidae Gilmore, 1924
Sinovenator changii gen. et sp. nov.

Holotype. IVPP (Institute of Vertebrate Paleontology and Paleoanthropology, Beijing) V 12615, a disarticulated partial skull and skeleton.

Referred specimen. IVPP V12583, an incomplete, articulated postcranial skeleton.

Etymology. 'Sinae', Latin referring to China, plus 'Venator', Latin for hunter. The specific name honours Meeman Chang of the IVPP for her significant role in the study of the Jehol fauna.

Locality and horizon. Lujiatun and Yanzigou, Shanyuan, western Liaoning, China; lowest part of Yixian Formation, older than 128 Myr (ref. 14; Hauterivian?); associated vertebrate fossils include the ceratopsian *Psittacosaurus*, the primitive ornithischian *Jeholosaurus*¹⁵, and the primitive mammal *Repenomamus*¹⁶.

Diagnosis. A troodontid with the following derived characters: straight and vertical anterior margin of antorbital fenestra; frontal with a vertical lamina bordering the lacrimal; surangular T-shaped

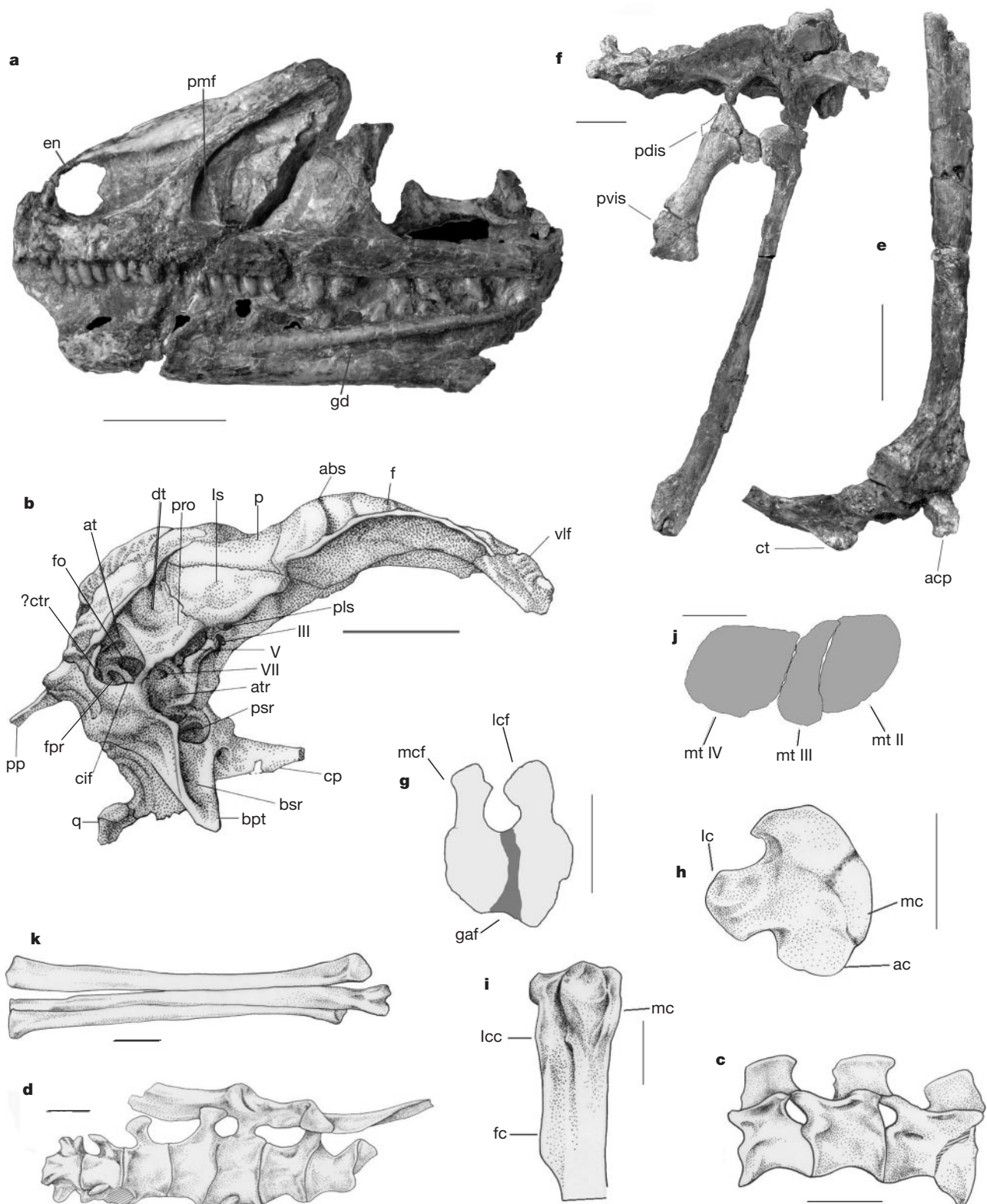


Figure 1 Select elements of *Sinovenator changii*. **a**, Right lateral view of the rostrum (IVPP V 12615). **b**, Right lateral view of the braincase (IVPP V 12615). **c**, Three anterior dorsals in lateral view (IVPP V 12615). **d**, sacrum in ventral view (IVPP V 12615). **e**, Right scapula and coracoid in lateral view (IVPP V 12615). **f**, Right pelvis in lateral view (IVPP V 12615); the presence of posterodorsal and posteroventral processes of the ischium is based on IVPP V 12583. **g**, Right femur in distal view (IVPP V 12615). **h**, **i**, Right tibia in proximal (**h**) and anterior (**i**) views (IVPP V 12583). **j**, **k**, Left metatarsus in proximal (**j**) and anterior (**k**) views (IVPP V 12583). **abs**, anterior boundary of the supratemporal fossa; **ac**, anterior condyle; **acp**, acromion process; **at**, accessory tympanic recess; **atr**, anterior tympanic recess; **bpt**, basipterygoid process; **bsr**, basipterygoid pneumatic recess; **cif**, crista

interfenestralis; **ct**, coracoid tubercle; **ctr**, caudal tympanic recess; **cp**, cultriform process; **dt**, dorsal tympanic recess; **en**, external naris; **f**, frontal; **fc**, fibular crest; **fo**, fenestra ovalis; **fpr**, fenestra pseudorotundum; **gaf**, extensor groove on femur; **gd**, groove on dentary; **III**, the third cranial nerve; **lc**, lateral condyle; **lcc**, lateral cnemial crest; **lcf**, lateral condyle of femur; **ls**, laterosphenoid; **mc**, medial condyle; **mcf**, medial condyle of femur; **mt II–IV**, metatarsal II–IV; **pdis**, posterodorsal process of ischium; **pls**, pit on laterosphenoid; **pmf**, promaxillary fenestra; **p**, parietal; **pp**, paroccipital process; **pro**, prootic; **psr**, parasphenoid pneumatic recess; **pvis**, posteroventral process of ischium; **q**, quadrate; **V**, fifth cranial nerve exit; **VII**, seventh cranial nerve exit; **vlf**, vertical laminae of frontal.

in cross-section; and a prominent lateral cnemial crest continuous with the fibular crest.

Description and comparison. *Sinovenator* is a small theropod, with an estimated length of less than 1 m. The skull is triangular in lateral view, with a shallow snout and a deep dorsally convex postorbital region. The supratemporal fenestra is reduced in dorsal view. The exposed endocast of *Sinovenator* is similar to the one reconstructed for *Archaeopteryx*¹⁷.

The external naris is significantly larger than the premaxilla as in other troodontids (Fig. 1a). As in dromaeosaurids¹⁸ and most theropods, but not other troodontids¹⁹, there is a long and slender posterior process of the premaxilla, which excludes the maxilla from the external naris. The sharp-rimmed antorbital fossa contains an antorbital fenestra, a maxillary fenestra and a promaxillary fenestra (Fig. 1a), the last one of which is not present in other troodontids. Unusually, the lateral edge of the nasal is deflected and faces dorsally rather than laterally as in most other theropods. It forms a small shelf adjacent to the nasomaxillary suture. The jugal is similar to that of dromaeosaurids in that the postorbital process is long, slender and oriented posteriorly. A pneumatic fossa is present on the posterior surface of the quadrate shaft, as in some other troodontid taxa²⁰.

The braincase shows many troodontid features including an oval foramen magnum, which is taller than it is wide, a shallow, 'V'-shaped incisure between the basal tubera, absence of a basisphenoid recess, and a pit on the ventral surface of the laterosphenoid

(Fig. 1b). Periotic pneumatic systems present in *Sinovenator* include the anterior tympanic recess, an accessory tympanic recess dorsal to the crista interfenestralis as in *Archaeopteryx*, a dorsal tympanic recess on the prootic, and a small caudal tympanic recess (Fig. 1b) that opens into the columellar recess, closely resembling those of *Archaeopteryx*^{21,22} and *Shuvuua*²³. *Sinovenator* lacks the well-developed subotic recess seen in other troodontids.

The basiptyergoid processes taper distally and bear prominent basiptyergoid recesses on their anterodorsal surfaces. A large fossa is adjacent to the base of the cultriform process (the parasphenoid recess) and is also observed in the basal dromaeosaurid *Sinornithosaurus*¹¹. This fossa is continuous with the anterior tympanic recess owing to the lack of a prominent otosphenoidal crest that delimits the 'lateral depression' of most other troodontids^{1,24}. In *Sinovenator* the parabasisphenoid rostrum is not bulbous. *Sinovenator* possesses several characters that are unique among troodontids including an accessory rod-like ventral process on the posterior pterygoid, a deep posterior end of the surangular, and a foramen magnum that is much larger than the occipital condyle.

As in other troodontids, a row of foramina is located in a groove posteriorly on the subtriangular labial surface of the dentary (Fig. 1a). The surangular is 'T'-shaped in cross-section, and the medial edge of the dorsal margin is significantly lower than the lateral edge and roofs over the internal mandibular fossa. The surangular is 'L'-shaped in cross-section in most other theropods.

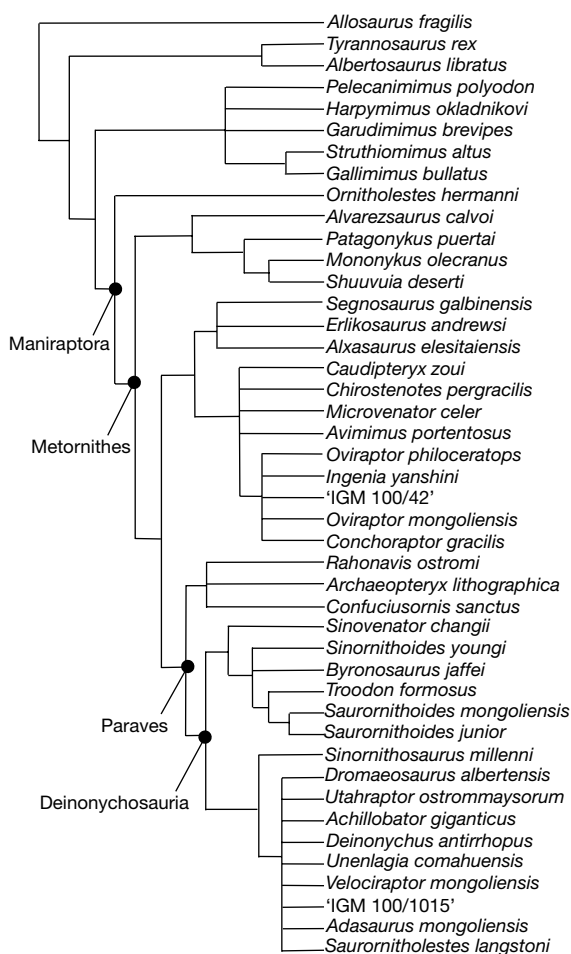


Figure 2 A proposed coelurosaurian phylogeny that is based on the strict consensus of the 336 equally most parsimonious trees (tree length, 576; consistency index, 0.43; retention index, 0.70). Examining the character distributions along the maniraptoran lineage indicates that some major character transformations, such as basally constricted

teeth, an 'L' shaped shoulder girdle and an opisthopubic pelvis, are characteristic of more inclusive maniraptoran clades and did not evolve at the base of or within the avialan clade.

The dentary teeth are closely packed rostrally as in other troodontids¹⁹, but the posterior maxillary teeth are slightly inset medially—an unusual condition in theropods. Denticles seem to be absent on the anterior teeth, and are small on the maxillary teeth as in dromaeosaurids.

The dorsals have moderately tall, fan-shaped neural spines; weakly developed hypapophyses are present on the anterior dorsal centra (Fig. 1c). Unlike other troodontids, which possess six sacralis, *Sinovenator* has five sacralis (Fig. 1d). The middle sacralis are wider than the anterior and posterior ones, as in the basal dromaeosaurid *Microraptor*²⁵ and avians (Fig. 1d).

The scapula and coracoid are very similar to those in dromaeosaurids and other derived maniraptorans^{11,26} in that the glenoid fossa faces laterally and the scapula and coracoid form an angle of less than 90° in lateral view (Fig. 1e). The ilium is small, being less than 60% the femoral length as in *Archaeopteryx* and the basal dromaeosaurid *Sinornithosaurus*. The pubic peduncle is deep and wide, and slightly posteriorly oriented as in *Archaeopteryx* and some dromaeosaurids. Unlike other troodontids^{4–6}, the pubis is posteriorly oriented. The ischium is similar to some advanced maniraptorans, including basal birds, in its small size, the distally positioned obturator process, and the presence of posterodorsal and posteroventral processes (Fig. 1f).

The femur has a derived distal end with a shallow extensor groove on the anterior surface, and a deep popliteal fossa that is constricted between the expanded distal ends of the medial and lateral condyles (Fig. 1g). The tibia is avian-like with an anteroposteriorly elongate proximal end, an incipient development of a medial cnemial crest (Fig. 1h), and a subrectangular distal end instead of the subtriangular one that is seen in most other theropods. Unusually, a prominent lateral cnemial crest continues onto the fibular crest (Fig. 1i). The metatarsus is very similar to the basal dromaeosaurids *Sinornithosaurus* and *Microraptor* except that the metatarsal II is markedly short and slender as in other troodontids^{25,27} (Fig. 1k). The proximal end of metatarsal III is visible anteriorly (Fig. 1j, k); however, the foot resembles more closely the subarctometatarsalian condition seen in the basal dromaeosaurid *Sinornithosaurus* than the non-arctometatarsalian condition.

Discussion. *Sinovenator* is from the lowermost, fluvial deposits of the Yixian Formation, and predates other known coelurosaurs with preserved feather-like integumentary structures from the lacustrine beds that occur higher within the formation. The facies that yielded the specimens of *Sinovenator* do not preserve soft tissue. At present, it is the oldest known taxon that can be referred confidently to the Troodontidae.

A comprehensive phylogenetic analysis (Supplementary Information) shows that *Sinovenator* is the most basal troodontid found so far, and that Troodontidae is the sister group to Dromaeosauridae within a monophyletic Deinonychosauria, which in turn is the sister group of Avialae (Fig. 2), as proposed in some previous studies^{6,10,12}.

The discovery of *Sinovenator* is significant in that, as the most primitive member of the Troodontidae, it resets character polarities at the base of this clade. The absence of a bulbous parasphenoid capsule in *Sinovenator* removes much of the support for a monophyletic Ornithomimosaur–Troodontid clade, Bullatosauria^{1,8,28,29}, and our phylogenetic analysis indicates that this character evolved independently in the two lineages. Exclusion of *Sinovenator* from the phylogenetic analysis posits troodontids as the sister taxon to an Oviraptorosaur–Therizinosaur clade¹³, mainly owing to the numerous teeth with large denticles, hollow basipterygoid processes, and absence of a tertiary antorbital opening shared by higher troodontids and therizinosaurs. Again, the absence of these derived characters in *Sinovenator* indicates that these similarities were acquired convergently, and underscores the pivotal role of *Sinovenator* in reconstructing maniraptoriform phylogeny. Gaining a full understanding of character evolution within Coelurosauria may not be possible if taxon sampling is carried out at higher taxonomic levels,

and the benefits of sampling species outweighs the problems caused by missing data in some fragmentary taxa.

The discovery of *Sinovenator changii* improves our understanding of the transition of several salient osteological characters to birds, notably, the laterally projected glenoid of the scapulacoracoid and the opisthopubic pelvis. In maniraptorans, a laterally projected glenoid of the scapulacoracoid can now be considered as diagnostic of the Deinonychosaur–Avialan clade (Paraves), and the opisthopubic pelvis as diagnostic of an even more inclusive clade (Metornithes) rather than as homoplastic in dromaeosaurids and avialans as suggested previously^{6,9,10,12}. Notably, small body size seems to be primitive for Paraves, and the trend toward miniaturization in early bird evolution¹² is a continuation of the paravian trend²⁵. Unexpectedly, this trend is reversed independently in dromaeosaurids and troodontids toward an increasing body size, and this is accompanied by homoplastic reversals in some key characters, such as the propubic or near-propubic pelves of derived troodontids and the bizarre dromaeosaurid *Achillobator*³⁰. Convergent evolution of other features that have caused confusion about the phylogenetic position of troodontids, such as the large denticles and pneumatic bullae, also occurred late in the evolution of this group in a subclade of larger-bodied troodontid taxa. □

Received 10 July; accepted 26 November 2001.

- Currie, P. J. Cranial anatomy of *Stenonychosaurus inequalis* (Saurischia, Theropoda) and its bearing on the origin of birds. *Can. J. Earth Sci.* **22**, 1643–1658 (1985).
- Currie, P. J. Bird-like characteristics of the jaws and teeth of troodontid theropods (Dinosauria: Saurischia). *J. Vert. Paleontol.* **7**, 72–81 (1987).
- Currie, P. J. & Zhao, X.-J. A new troodontid (Dinosauria, Theropoda) braincase from the Dinosaur Park Formation (Companionian) of Alberta. *Can. J. Earth Sci.* **30**, 2224–2230 (1993).
- Osmolska, H. & Barsbold, R. in *The Dinosauria* (eds Weishampel, D. B., Dodson, P. & Osmolska, H.) 259–268 (Univ. California Press, Berkeley, 1990).
- Russell, D. A. & Dong, Z.-M. A nearly complete skeleton of a new troodontid dinosaur from the Early Cretaceous of the Ordos Basin, Inner Mongolia, People's Republic of China. *Can. J. Earth Sci.* **30**, 2163–2173 (1993).
- Gauthier, J. A. Saurischian monophyly and the origin of birds. *Mem. Calif. Acad. Sci.* **8**, 1–55 (1986).
- Russell, D. A. & Dong, Z. The affinities of a new theropod from the Alxa Desert, Inner Mongolia, China. *Can. J. Earth Sci.* **30**, 2107–2127 (1993).
- Holtz, T. R. Jr The phylogenetic position of the Tyrannosauridae: implications for theropod systematics. *J. Paleontol.* **68**, 1100–1117 (1994).
- Forster, C. A., Sampson, S. D., Chiappe, L. M. & Krause, D. W. The theropod ancestry of birds: new evidence from the Late Cretaceous of Madagascar. *Science* **279**, 1915–1919 (1998).
- Makovicky, P. & Sues, H.-D. Anatomy and phylogenetic relationships of the theropod dinosaur *Microvenator celer* from the Lower Cretaceous of Montana. *Am. Mus. Novitates* **3240**, 1–27 (1998).
- Xu, X., Wang, X.-L. & Wu, X.-C. A dromaeosaurid dinosaur with a filamentous integument from the Yixian Formation of China. *Nature* **401**, 262–266 (1999).
- Sereno, P. C. The evolution of dinosaurs. *Science* **284**, 2137–2147 (1999).
- Norell, M. A., Clark, J. M. & Makovicky, P. in *New Perspectives on the Origin and Early Evolution of Birds* (eds Gauthier, J. & Gall, L. F.) 49–67 (Peabody Museum of Natural History, Yale Univ., Newhaven, 2001).
- Wang, X.-L. et al. Stratigraphic sequence and vertebrate bearing beds of the lower part of the Yixian Formation in Sihetun and neighbouring area, western Liaoning, China. *Vert. Palasiatica* **36**, 81–101 (1998).
- Xu, X., Wang, X.-L. & You, H.-L. A primitive ornithomimid from the Early Cretaceous Yixian Formation of Liaoning. *Vert. Palasiatica* **38**, 318–325 (2000).
- Li, J.-L., Wang, Y., Wang, Y.-Q. & Li, C.-K. A new family of primitive mammal from the Mesozoic of western Liaoning, China. *Chinese Sci. Bull. (Engl.)* **46**, 782–786 (2001).
- Chatterjee, S. *The Rise of Birds* (John Hopkins Univ. Press, Baltimore, 1997).
- Barsbold, R. & Osmolska, H. The skull of *Velociraptor* (Theropoda) from the late Cretaceous of Mongolia. *Acta Palaeontol. Polon.* **44**, 189–219 (1999).
- Norell, M. A., Makovicky, P. & Clark, J. M. A new troodontid theropod from Ukhaa Tolgod, Mongolia. *J. Vert. Paleontol.* **20**, 7–11 (2000).
- Barsbold, R., Osmolska, H. & Kurzanov, S. On a new troodontid (Dinosauria: Theropoda) from the Early Cretaceous of Mongolia. *Acta Palaeontol. Polon.* **32**, 121–132 (1997).
- Walker, A. D. in *The Beginnings of Birds* (eds Hecht, M. K., Ostrom, J. H., Viohl, H. & Wellnhofer, P.) 123–134 (Freunde des Jura-Museums, Eichstatt, 1985).
- Witmer, L. M. The craniofacial air sac system of Mesozoic birds (Aves). *Zool. J. Linn. Soc.* **100**, 327–378 (1990).
- Chiappe, L. M., Norell, M. A. & Clark, J. M. The skull of a relative of the stem-group bird *Mononykus*. *Nature* **392**, 275–278 (1998).
- Barsbold, R. Saurornithomimididae, a new family of small theropod dinosaurs from Central Asia and North America. *Palaeontol. Polon.* **30**, 5–22 (1974).
- Xu, X., Zhou, Z.-H. & Wang, X.-L. The smallest known non-avian theropod dinosaur. *Nature* **408**, 705–708 (2000).
- Norell, M. A. & Makovicky, P. J. Important features of the dromaeosaurid skeleton II: information from newly collected specimens of *Velociraptor mongoliensis*. *Am. Mus. Novitates* **3282**, 1–45 (1999).
- Xu, X. & Wang, X.-L. Troodontid-like pes in the dromaeosaurid *Sinornithosaurus*. (Spec. Publ.) 179–188 (Paleontology Society of Korea, Seoul, 2000).

28. Kurzanov, S. M. Braincase structure in the carnosaur *Itemirus* n. gen. and some aspects of the cranial anatomy of dinosaurs. *Paleontol. Zh.* **1976**, 127–137 (1976).
29. Holtz, T. R. Jr Phylogenetic taxonomy of the Coelurosauria (Dinosauria: Theropoda). *J. Paleontol.* **70**, 536–538.
30. Perle, A., Norell, M. & Clark, J. A new maniraptoran theropod *Achillobator giganticus* (Dromaeosauridae) from the Upper Cretaceous of Burkhan, Mongolia. (Dept Geol., Natl Univ. Mongolia, Ulaanbaatar, 1999).

Supplementary Information accompanies the paper on *Nature's* website (<http://www.nature.com>).

Acknowledgements

We thank Z.-H. Zhou for help; H.-J. Wang and B. Long for specimen preparation; R.-S. Li for drawings; M. Ellison for photographs; and members of the Liaoxi expedition team of the IVP. This work was supported by grants from the Special Funds for Major State Basic Research Projects of China, the National Geographic Society, the Chinese Natural Science Foundation and the Chinese Academy of Sciences. M.A.N. and P.J.M. are supported by the American Museum of Natural History and the Field Museum, respectively.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to X.X. (e-mail: xxu@midwest.com.cn) or M.A.N. (e-mail: norell@amnh.org).

Chameleon radiation by oceanic dispersal

C. J. Raxworthy*, M. R. J. Forstner† & R. A. Nussbaum‡

* American Museum of Natural History, Central Park West at 79th Street, New York, New York 10024-5192, USA

† Department of Biology, Southwest Texas State University, San Marcos, Texas 78666, USA

‡ Museum of Zoology, University of Michigan, Ann Arbor, Michigan 48109-1079, USA

Historical biogeography is dominated by vicariance methods that search for a congruent pattern of fragmentation of ancestral distributions produced by shared Earth history^{1–3}. A focus of vicariant studies has been austral area relationships and the break-up of the supercontinent Gondwana^{3–5}. Chameleons are one of the few extant terrestrial vertebrates thought to have biogeographic patterns that are congruent with the Gondwanan break-up of Madagascar and Africa^{6,7}. Here we show, using molecular and morphological evidence for 52 chameleon taxa, support for a phylogeny and area cladogram that does not fit a simple vicariant history. Oceanic dispersal—not Gondwanan break-up—facilitated species radiation, and the most parsimonious biogeographic hypothesis supports a Madagascan origin for chameleons, with multiple ‘out-of-Madagascar’ dispersal events to Africa, the Seychelles, the Comoros archipelago, and possibly Reunion Island. Although dispersal is evident in other Indian Ocean terrestrial animal groups^{8–16}, our study finds substantial out-of-Madagascar species radiation, and further highlights the importance of oceanic dispersal as a potential precursor for speciation.

The break-up of Gondwana provides one of the most obvious test cases for vicariance biogeography, with the separation of Africa, Madagascar, the Seychelles and India providing a well documented geologic history⁴. Magnetic anomaly data and Mesozoic timescales reveal that Madagascar and Greater India first broke away from Africa 165 Myr ago, with movement ending by 121 Myr ago¹⁷; Greater India and Madagascar separated approximately 88 Myr ago based on dating of Marion hot-spot-related volcanism¹⁸; and the

Seychelles granitic archipelago split from India during formation of the Deccan basalt province between 69 and 65 Myr ago¹⁹. The main source of current controversy concerns the sequence and timing of separation of Antarctica with India, South America and Australia between 80–100 Myr ago^{18,20}.

Gondwanan origins have been proposed for extant terrestrial vertebrates in Madagascar⁸; however, no modern systematic evidence congruent with a Gondwanan break-up vicariant pattern has been reported to date. Chameleons (family Chamaeleonidae: subfamily Chamaeleoninae²¹) are claimed to be an ancient Gondwanan group, on the basis of limited immunological distances and a calibrated albumin molecular clock⁷. Previous cladistic studies, using 11–24 morphological characters, supported conflicting biogeographic hypotheses that are partly congruent with Gondwanan break-up^{6,22}, or that suggest a post-Gondwanan, Madagascan origin for chameleons²³.

To test the Gondwanan vicariance hypothesis, we conducted a phylogenetic analysis that included 52 chameleon taxa (approximately 40% of all chameleon species). These taxa represent all Chamaeleoninae genera and subgenera, and include species found in each continental region (including the granitic Seychelles), and the Comoros archipelago and Reunion Island (both of volcanic origin). To test the monophyly of chameleons, the two other Chamaeleonidae subfamilies (Leiolepidinae and Agaminae) were also included. Molecular characters (mitochondrial DNA (mtDNA), NADH subunit 4 and adjacent transfer RNA sequences), morphological characters, and behavioural/life-history characters were used to produce a combined data set of 644 potentially parsimony-informative characters.

Three equally parsimonious trees were found for the combined equally weighted data set. These results (Fig. 1) support the monophyly of chameleons, with the two most basal clades representing dwarf chameleons (*Brookesia*), which are endemic to Madagascar. Basal to the crown clade of ‘typical’ chameleons (*Furcifer*, *Calumma* and *Chamaeleo*) are the endemic African genera, *Rhampholeon* and *Bradypodion*. The resulting area cladogram (Fig. 1) has the following features: (1) The two most basal lineages are distributed in Madagascar; (2) The more derived lineages occur in Madagascar, Africa, the Seychelles and India, with the two most basal lineages of this clade occurring in Africa; (3) the India lineage is sister to an African clade; (4) the Seychelles lineage is sister to a Madagascar clade; and (5) the Comoros lineage is sister to a Madagascan clade.

The corresponding chameleon continental area cladogram (Fig. 2a) cannot be reconciled²⁴ with proposed Gondwanan break-up

Table 1 Chameleon area cladogram reconciled with origin hypotheses

Biogeographic hypothesis	Minimum assumption*		
	Extinction†	Dispersal	Total
Gondwanan origin 1‡			
Vicariance only	38	0	38
Vicariance and dispersal	3	5	8
Gondwanan origin 2§			
Vicariance only	25	0	25
Vicariance and dispersal	1	5	6
Partial Gondwanan origin			
Madagascar + Seychelles + India	0	12	12
Madagascar + Seychelles + India	1	11	12
Seychelles + India	0	14	14
Post-Gondwanan origin			
Madagascar	0	5	5
Africa	0	6	6
Seychelles	0	12	12
India	0	11	11

* Minimum number of assumption events to reconcile area cladogram with biogeographic hypothesis.

† Minimum number of losses²⁴. Inclusion of all missing nodes (items of error²¹) would add additional extinction events.

‡ Conventional hypothesis.

§ Hypothesis of ref. 20.

vicariance models (Fig. 2b, c) or models of post-Gondwanan origin, without invoking assumptions of dispersal or extinction. The most parsimonious biogeographic hypothesis that minimizes dispersal or extinction events supports a post-Gondwanan origin in Madagascar, with subsequent oceanic dispersal from Madagascar. These continental dispersal events include three dispersals to Africa, and

a single dispersal to both India and the Seychelles (Table 1). Other less parsimonious biogeographic hypotheses place chameleon origins in Madagascar, Africa or Gondwana, and also include a minimum of five dispersal events (Table 1). An utter rejection of oceanic dispersal requires a minimum of 25 or 38 independent extinction events, depending on the choice of Gondwanan break-up model (Table 1). Both of these 'no-dispersal' hypotheses reject a role for tectonic vicariance facilitating the origin of the major species radiations, because all generic lineages would have to predate Gondwanan break-up.

Previously proposed alternative phylogenies (and their associated biogeographic patterns) are substantially less parsimonious compared with our hypothesis, requiring greater numbers of character state changes. Forcing a strict Gondwanan break-up topology (Fig. 2b, c) on the chameleon area clades (AF(M(SE,I))); see Fig. 2 legend for definitions) increases tree length by a minimum of 49 steps. The previously proposed Gondwanan vicariance split for Brookesini (*Brookesia*/*Rhampholeon*) and Chamaeleonini (*Bradypodion* and *Chamaeleo*/*Calumma* and *Furcifer*)⁶ requires a minimum of 25 additional steps. A Madagascan or African origin hypothesis with a minimum number of three dispersal events (to the Seychelles, India, and either Africa or Madagascar) requires a minimum of 18 and 15 additional steps, respectively.

Evidence corroborating oceanic dispersal for chameleons is provided by the sister species *Furcifer cephalolepis* and *F. polleni*, which are endemic to the volcanic Comoros islands, Moheli and

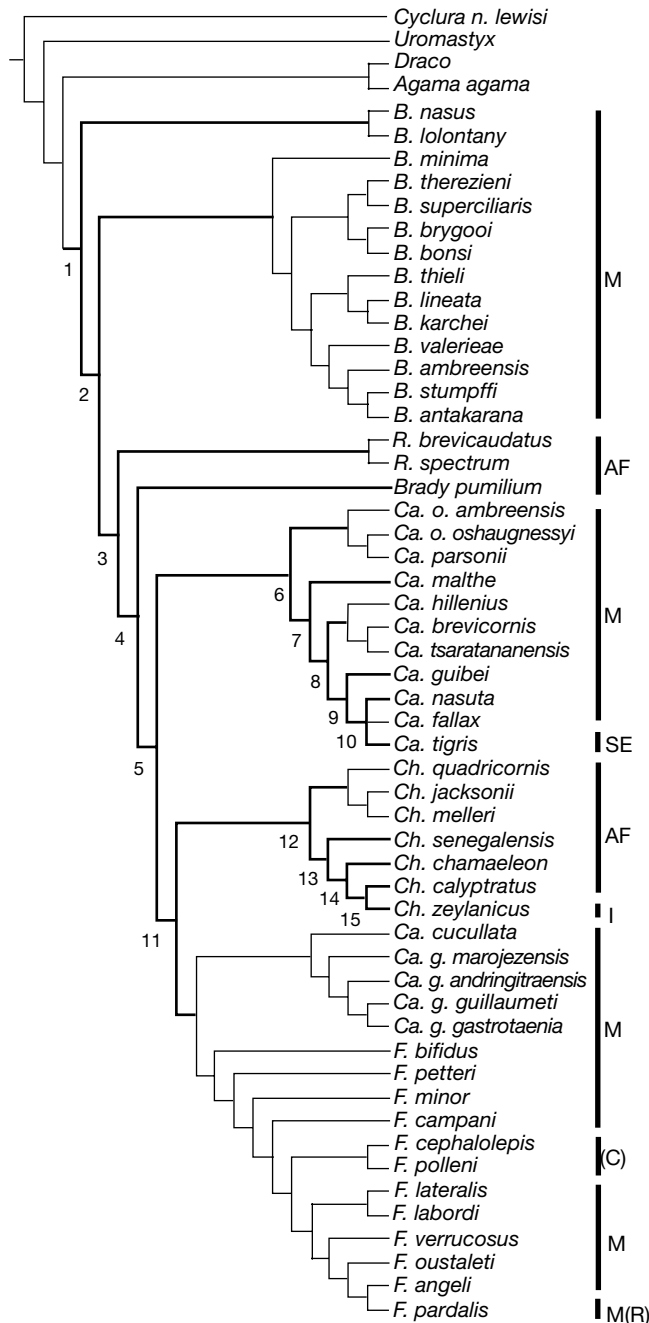


Figure 1 Inter-relationships of chameleons and corresponding area distributions based on equal-weighted molecular and morphological data and maximum parsimony. *B.*, *Brookesia*; *R.*, *Rhampholeon*; *Brady.*, *Bradypodion*; *Ca.*, *Calumma*; *Ch.*, *Chamaeleo*; *F.*, *Furcifer*. Taxon distribution: AF, Africa, Arabian Peninsula and Near-Middle East; M, Madagascar; I, India, Pakistan and Sri Lanka; SE, Seychelles; (C), Comoros; (R), Reunion. Non-continental (volcanic origin) areas are shown in brackets. Continental area cladogram is indicated in bold, with percentage bootstraps (all data/all data except third positions, with successive weighting): 1, 100/100; 2, 73/83; 3, 77/92; 4, 100/100; 5, 76/86; 6, 54/69; 7, 99/100; 8, 81/66; 9, 98/98; 10, 53/32; 11, 78/71; 12, 100/100; 13, 100/100; 14, 100/100; 15, 100/100.

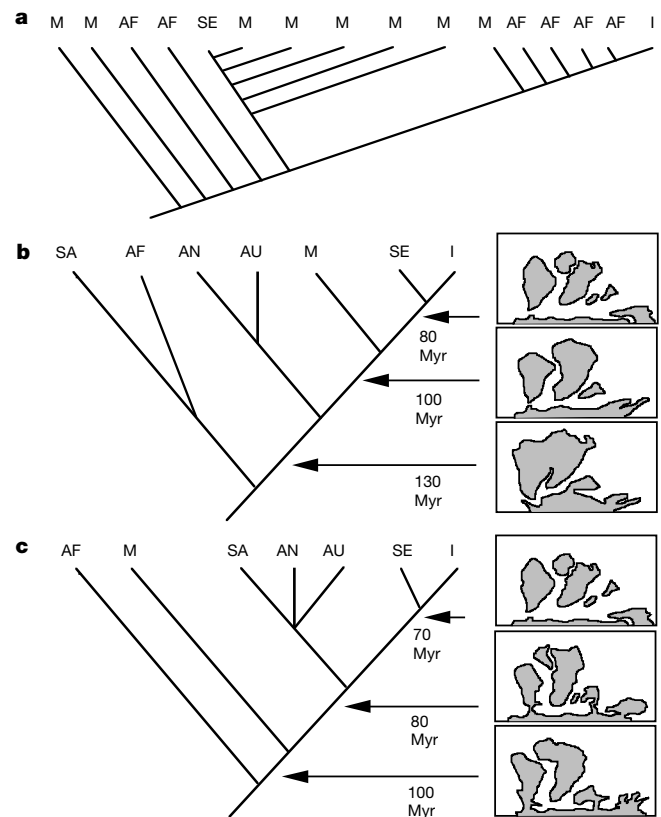


Figure 2 Continental area cladograms for the Indian Ocean region based on chameleon phylogeny and geological break-up models. AF, Africa; AN, Antarctica; AU, Australia; M, Madagascar; SA, South America; SE, the Seychelles; I, India and Sri Lanka. **a**, Chameleon area cladogram based on Fig. 1. **b**, Conventional hypothesis of Gondwanan break-up for the Indian Ocean region^{17–19}. **c**, Alternative hypothesis of Gondwanan break-up for the Indian Ocean region proposed by ref. 20. The chameleon area cladogram is incongruent with both of the break-up hypotheses.

Grand Comoro, respectively⁸. These islands were formed between 0.13 and 5.4 Myr ago²⁵, and never had contact with other land-masses. The direct ancestor of these endemic chameleons could only have reached the archipelago by means of oceanic dispersal. Our results find these sister species most closely related to the *F. oustaleti* and *F. lateralis* groups in Madagascar, which are separated by 300 km of ocean from the Comoros archipelago. The distribution of the chameleon, *Furcifer pardalis*, in both Madagascar and Reunion Island, may represent another natural oceanic dispersal, as differences in dorsal crest and hemipene morphology found between these island populations do not suggest a modern human introduction, subsequent to human occupation of Reunion Island, in the seventeenth century.

Other evidence is not inconsistent with a post-Gondwanan origin for chameleons. The oldest known chameleon fossils are no older than the Miocene epoch (*Rhampholeon* from Kenya, 18 Myr ago²⁶; *Chamaeleo intermedius* from Kenya, 16–18 Myr ago²⁷; and *Chamaeleo caroli* from Central Europe, 26 Myr ago²⁸). The earliest known fossil acrodont lizard *Mimeosaurus* (Upper Cretaceous period)²¹ is also younger than the Africa–Madagascar separation (extant acrodonts are confined to the Agaminae, Leiolepidinae and Chamaeleoninae). However, all extant lizard groups are poorly represented in the fossil record. Molecular clock mtDNA divergences (assuming a vertebrate ectotherm mtDNA divergence rate of 0.4–0.6% per Myr ago^{11,29}) are also consistent with a post-Gondwanan origin for chameleons. Sequence divergence between chameleons and their sister group (*Agama*, *Draco*) is 28–36%, giving an estimated divergence time of 47–90 Myr, contemporary or younger than the separation of Madagascar and India. Sequence divergence between the basal *Brookesia* clade and all other chameleons is 21–27% (35–68 Myr), *Rhampholeon* and all typical chameleons 20–26% (33–65 Myr), and *Chamaeleo* and *Furcifer* 17–23% (28–58 Myr).

Despite the Gondwanan vicariant history of Madagascar and the Indian Ocean, many regional species radiations (ants, spiders, tortoises, geckos, skinks, rodents, primates) are reported to show transmarine migrations, and have post-Gondwanan origins in Africa or Asia^{8–16}. Our results provide the strongest support for a post-Gondwanan species radiation originating in Madagascar, with subsequent out-of-Madagascar dispersal to Africa, the Seychelles, the Comoros, and possibly Reunion Island. Invasions into Africa from Madagascar are currently not well documented, although dispersal through Madagascar, from Asia to Africa, has been concluded recently for rodents¹⁵. Surprisingly, the absence of biogeographic patterns congruent, or largely congruent with Gondwanan break-up in the Indian Ocean (and other austral regions), has thus far received little attention, although two poorly supported geological hypotheses (Pacifica supercontinent and expanding earth) have been proposed to maintain the plausibility of alternative vicariant patterns^{3,5}. Our analyses provide evidence for considerable oceanic dispersal by chameleons, and support the hypothesis that dispersal, rather than continental break-up, was the precursor for species radiation. □

Methods

For phylogenetic analysis, a total of 85 morphological and behavioural characters, and 972 bases of mtDNA (236 3' terminal codons of ND4, tRNA^{Ser}, tRNA^{His} and tRNA^{Leu}) were used. Sequence alignment was made using CLUSTAL W (see <http://www.mbio.ncsu.edu/BioEdit/bioedit.html>) under a wide variety of gap penalty assignments, and refined by eye. For the list of morphological and behavioural characters see Supplementary Information. The non-Chamaeleonidae specified outgroup taxa was *Cyclura* (Iguanidae). The 55 in-group Chamaeleonidae (from ref. 21) taxa include Leiolepidinae (*Uromastix*), Agaminae (*Agama* and *Draco*) and Chamaeleoninae (14 *Brookesia*, 2 *Rhampholeon*, 1 *Bradypodion*, 16 *Calumma*, 7 *Chamaeleo* and 12 *Furcifer*). Chameleon taxa are shown in Fig. 1. We collected all species from Madagascar and the Seychelles; we obtained other species from tissue and museum collections or commercial sources. Phylogenetic analysis was performed using parsimony with PAUP version 4.0b2. All characters were initially given equal weight and all characters were unordered except eight morphological characters, which were ordered based on intermediacy. Gaps were treated as a fifth base. Searches were heuristic with 10 replicate searches with random stepwise addition of taxa. No upper limit

was imposed on the maximum number of trees saved. Successive weighting was based on maximum rescaled consistency index (RC). Bootstrap branch support was based on 100 bootstrap replicates (each replicate with 10 replicate heuristic searches with random stepwise addition of taxa). MacClade 4.0 was used to determine step costs for alternative phylogenetic hypotheses.

Equal weighting of the combined molecular and morphological characters (644 parsimony informative) recovered three maximum parsimony trees (5,096 steps, consistency index, CI = 0.281, RC = 0.143, strict consensus; Fig. 1). Relaxing parsimony by one step increased the total tree number to 35, with the 50% majority rule consensus congruent with the most parsimonious strict consensus cladogram, but the strict consensus showing basal polytomies. Successive weighting recovered the same strict consensus and area cladogram found with equal weighting (3 trees, 7,016 steps, CI = 0.529, RC = 0.374). Reverse successive weighting, used to explore potential secondary phylogenetic signal³⁰, recovered 262 trees and a strict consensus tree showing: (1) almost no resolution (except within *Brookesia* and *Calumma*); (2) chameleons paraphyletic (with respect to *Draco*); and (3) third position molecular characters representing most of the high reverse-weighted characters. Deletion of third positions, and otherwise using equal weighting, resulted in a strict consensus tree almost identical to the entire data set (incongruence confined to minor polytomies within *Furcifer*) and an area cladogram completely congruent with Fig. 2a. (84 trees, 2,670 steps, CI = 0.336, RC = 0.194). Successive weighting and excluding third positions resulted in a strict consensus area cladogram completely congruent with Fig. 2a (12 trees, 5,141 steps, CI = 0.607, RC = 0.466).

Received 15 October; accepted 19 November 2001.

- Nelson, G. & Platnick, N. I. *Systematics and Biogeography: Cladistics and Vicariance* (Columbia University, New York, 1981).
- Morrone, J. J. & Crisci, J. V. Historical biogeography: introduction to methods. *Annu. Rev. Ecol. Syst.* **26**, 373–401 (1995).
- Humphries, C. J. & Parenti, L. R. *Cladistic Biogeography* 2nd edn (Oxford Univ. Press, Oxford, 1999).
- Brown, J. H. & Lomolino, M. V. *Biogeography* 2nd edn (Sinauer, Sunderland, 1998).
- Crisci, J. V., Cigliano, M. M., Morrone, J. J. & Roig-Juñent, S. Historical biogeography of southern South America. *Syst. Zool.* **40**, 152–171 (1991).
- Klaver, C. J. J. & Böhm, W. Phylogeny and classification of the chamaeleonidae (Suria) with special reference to hemipenis morphology. *Bonn. Zool. Mon.* **22**, 1–64 (1986).
- Hofman, A., Maxson, L. R. & Arntzen, W. Biochemical evidence pertaining to the taxonomic relationships within the family Chamaeleonidae. *Amphibia-Reptilia* **12**, 245–265 (1991).
- Blanc, C. P. in *Biogeography and Ecology in Madagascar* (eds Battistini, R. & Richard-Vindard, G.) 501–614 (Junk, The Hague, 1972).
- Arnold, E. N. Indian Ocean giant tortoises: their systematics and island adaptations. *Phil. Trans. R. Soc. Lond. B* **286**, 127–145 (1979).
- Kluge, A. G. & Nussbaum, R. A. A review of African–Madagascan gekkonid lizard phylogeny and biogeography (Squamata). *Misc. Publ. Mus. Zool. Univ. Michigan* **183**, 1–20 (1995).
- Coccone, A., Amato, G., Gratry, O. C., Behler, J. & Powell, J. R. A molecular phylogeny of four endangered Madagascar tortoises based on MtDNA sequences. *Mol. Phylog. Evol.* **12**, 1–9 (1999).
- Fisher, B. L. in *Biogeography of Madagascar* (ed. Lourenco, W. R.) 457–466 (Orstom, Paris, 1996).
- Yoder, A. D. in *Biogeography of Madagascar* (ed. Lourenco, W. R.) 245–258 (Orstom, Paris, 1996).
- Mausfeld, P., Vences, M., Schmitz, A. & Veith, M. First data on the molecular phylogeography of scincoid lizards of the genus *Mabuya*. *Mol. Phylog. Evol.* **17**, 11–14 (2000).
- Jansa, S. A., Goodman, S. M. & Tucker, P. K. Molecular phylogeny and biogeography of the native rodents of Madagascar (Muridae: Nesomyinae): a test of the single-origin hypothesis. *Cladistics* **15**, 253–270 (1999).
- Griswold, C. E. in *Diversity and Endemism in Madagascar* (eds Lourenco, W. R. & Goodman, S. M.) 345–354 (Société de Biogéographie, Paris, 2000).
- Rabinowitz, P. D., Coffin, M. F. & Falvey, D. The separation of Madagascar and Africa. *Science* **220**, 67–69 (1983).
- Storey, M. et al. Timing of hot spot-related volcanism and the breakup of Madagascar and India. *Science* **267**, 852–855 (1995).
- Storey, B. C. The role of mantle plumes in continental breakup: case histories from Gondwanaland. *Nature* **377**, 301–308 (1995).
- Hay, W. W. et al. in *Evolution of the Cretaceous Ocean–Climate System* (Spec. Pap. 332) (eds Barrera, E. & Johnson, C.) 1–48 (Geological Society of America, Boulder, 1999).
- Frost, D. R. & Etheridge, R. A phylogenetic analysis and taxonomy of the iguanian lizards (Reptilia: Squamata). *Univ. Kansas Mus. Nat. Hist. Misc. Publ.* **81**, 1–65 (1989).
- Rieppel, O. & Crumley, C. Paedomorphosis and skull structure in Malagasy chameleons (Reptilia: Chamaeleonidae). *J. Zool. Lond.* **243**, 351–380 (1997).
- Rieppel, O. The phylogenetic relationships within the Chamaeleonidae, with comments on some aspects of cladistic analysis. *Zool. J. Linn. Soc.* **89**, 41–62 (1987).
- Page, R. D. M. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Syst. Biol.* **43**, 58–77 (1994).
- Emerick, C. M. & Duncan, R. A. Age progressive volcanism in the Comores Archipelago, western Indian Ocean and implications for Somali plate tectonics. *Earth Planet. Sci. Lett.* **60**, 415–428 (1982).
- Rieppel, O., Walker, A. & Odhiambo, I. A preliminary report on a fossil Chamaeleonine (Reptilia: Chamaeleonidae) skull from the Miocene of Kenya. *J. Herpetol.* **26**, 77–80 (1992).
- Hillenius, D. A new chameleon from the Miocene of Fort Ternan, Kenya (Chamaeleonidae, Reptilia). *Beaufortia* **28**, 9–15 (1978).
- Moody, S. & Roček, Z. *Chamaeleo caroli* (Chamaeleonidae, Suria): a new species from the Lower Miocene of Central Europe. *Vestník Ústředního ústavu geologického* **55**, 85–92 (1980).
- Rand, D. M. Thermal habit, metabolic rate, and the evolution of mitochondrial DNA. *Trends Ecol. Evol.* **9**, 125–131 (1994).
- Trueman, J. W. H. Reverse successive weighting. *Syst. Biol.* **47**, 733–737 (1998).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com>).

Acknowledgements

We thank the authorities in Madagascar and the Seychelles for permission to conduct fieldwork. We thank J. Brown, T. Peterson, R. Prum, O. Rieppel, L. Trueb and E. Wiley for comments. C.J.R. and R.A.N. were supported by the National Geographic Society and the National Science Foundation, Washington.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.J.R. (e-mail: rax@amnh.edu). Sequences are deposited of GenBank under accession numbers AF443224–AF443275.

Antagonistic coevolution between the sexes in a group of insects

Göran Arnqvist*‡ & Locke Rowe*‡

* Department of Animal Ecology, Evolutionary Biology Centre, University of Uppsala, Norbyvägen 18d, SE-752 36 Uppsala, Sweden

† Department of Zoology, University of Toronto, Toronto, Ontario M5S 1A1, Canada

‡ The authors contributed equally to this work

In coevolutionary ‘arms races’ between the sexes, the outcome of antagonistic interactions may remain at an evolutionary standstill. The advantage gained by one sex, with any evolutionary exaggeration of arms, is expected to be matched by analogous counteradaptations in the other sex^{1,2}. This fundamental coevolutionary process may thus be hidden from the evolutionist’s eye^{3,4}, and no natural examples are known. We have studied the effects of male and female armament (clasping and anti-clasping morphologies) on the outcome of antagonistic mating interactions in 15 species of water strider, using a combination of experimental and phylogenetic comparative methods. Here we present, by assessing the independent effects of both species-specific level of arms escalation and small imbalances in the amounts of arms between the sexes within species, the consequences of a sexual arms race. Evolutionary change in the balance of armament between males and females, but not in the species-specific level of escalation, has resulted in evolutionary change in the outcome of sexually antagonistic interactions such as mating rate.

Evolutionary conflicts of interests between the sexes are ubiquitous^{1–4}. Such conflict is predicted to fuel sexually antagonistic coevolution^{1,2}, during which adaptations in one sex, which are harmful for individuals of the other sex, select for counteradaptations in the other sex to mitigate costs imposed by such adaptations. The resulting coevolution between the sexes is now recognized to be

a central process of evolution, with the potential to shape various interactions between the sexes^{2,5,6} and their gametes^{7–8}, as well as diversification⁹, speciation and extinction rates^{10–12}. At the core of this coevolutionary interaction is an arms race between the sexes that can include periods of both escalation and de-escalation of arms^{13–15}.

Theory suggests, however, that the outcome of antagonistic male–female interactions should remain relatively unchanged during an arms race because the build-up of arms in one sex may be balanced by a build-up in the other (Fig. 1a, 1–2). The consequences of such arms races on sexual interactions may thus be undetectable, which makes sexually antagonistic coevolution inherently difficult to show^{1–4}. Perhaps for this reason, we have no direct empirical evidence for a primary role of arms races in the evolution of sexual interactions in natural systems.

Despite an expectation of some evolutionary balance in the level of arms between the sexes, one sex may at least temporarily evolve a greater quantity of arms relative to the other (refs 13–15; and Fig. 1a, 3–4). In such cases, the evolutionary consequences of the arms race for interactions between the sexes may be exposed. For example, in an arms race in which it benefits males but not females to mate several times¹³, one may expect relatively high rates of mating in species where the advantage has shifted toward males (Fig. 1a, 4). The converse would be expected for species in which the advantage has shifted toward females (Fig. 1a, 3). Thus, by using tests that are based on phylogeny^{16,17}, one might uncover sexually antagonistic adaptations by analysing the effects on the change in sexual interactions caused by evolutionary change in the relative levels of arms between the sexes.

We have studied the coevolution of relative armament of the sexes and the outcome of sexually antagonistic interactions in 15 congeneric water strider species. Water striders (Heteroptera; Gerridae) are a group of semi-aquatic insects, which have become a model system in which to study sexually antagonistic coevolution. Several experimental in-depth studies, carried out on several different species and using different approaches (reviewed in refs 18, 19), have shown that there is intense and overt sexual conflict over mating rate that results from a strong asymmetry between the sexes in the relative costs and benefits of mating. Matings are preceded typically by a violent pre-mating struggle, in which females try to dislodge harassing males to avoid superfluous and costly mating^{18,19}.

The ability of males to withstand these struggles is related to various morphological grasping adaptations, such as exaggeration of prolonged clasping genitalia and a more flattened abdomen^{20,21}, which allow males to grasp females more firmly. Females’ ability to resist males is related to distinct morphological counteradaptations, such as prolongation of the female abdominal spines and the degree of downward tilting of the abdominal tip^{20,22}, which makes it more difficult for males to grasp females. We have shown elsewhere that the level of these arms in females coevolves closely with those in males within water striders²⁰. Thus, species can be ordinated along a

Table 1 Effects of morphological armament on the outcome of sexually antagonistic interactions

	Duration of pre-mating struggles			Male struggle success			Female mating activity			Female mating rate		
	β	t	P	β	t	P	β	t	P	β	t	P
Male persistence*	12.52	3.20	0.004	0.38	2.59	0.012	0.35	2.50	0.014	0.33	1.93	0.039
Female resistance*	–13.34	3.51	0.002	–0.47	3.28	0.003	–0.38	2.80	0.008	–0.35	2.09	0.029
Absolute level of arms (PC1)	0.83	0.87	0.399	0.06	1.71	0.113	0.02	0.73	0.482	0.02	0.40	0.693
Relative armament of the sexes (PC2)*	–14.83	3.13	0.004	–0.53	3.15	0.004	–0.45	2.83	0.007	–0.41	2.06	0.031

The effects of absolute and relative morphological armament on the outcome of sexually antagonistic interactions are shown. Multiple regression analyses, using behaviour as the dependent variable and either male persistence and female resistance, or PC1 and PC2, as independent variables. Regression models were based on phylogenetically independent contrasts ($n = 14$) and thus were forced through the origin¹⁹. Residuals did not differ significantly from normality (Kolmogorov–Smirnov tests, $P \geq 0.107$ for the first and $P \geq 0.101$ for the second group of models). Statistical power analyses of these models showed that our inability to detect any effects of absolute level of arms is not likely to be due to a lack of power. The combined probability of committing four type II errors was $\beta = 0.129$, assuming that absolute level of arms accounts for 10% of the variance in the dependent variable.

* P values reported test partial regression coefficients under directional null hypotheses (see Methods) and thus are one-tailed.

coevolutionary trajectory within a two-dimensional (2D) morphological space. These dimensions describe the degree of exaggeration in male arms that enhance persistence and in female arms that enhance resistance (Fig. 1b).

Such a trajectory is a predicted result of the coevolution of male and female arms^{13–15}. We determined whether the coevolution of male and female morphologies along and/or away from this trajectory leads to evolutionary changes in the outcome of antagonistic mating interactions. If the coevolutionary trajectory reflects a balance of arms, then evolutionary deviation away from the line should affect these interactions, whereas evolutionary movement along the line should have little effect. Specifically, we expect that as females become more armed relative to males, their ability to resist will be elevated and the duration and success of antagonistic interactions (that is, pre-mating struggles) will therefore decline, leading to reduced superfluous and costly mating among females.

We analysed the effects of evolution of morphological armament in the sexes on evolution of four behavioural traits: duration of pre-mating struggles, male struggle success, female mating activity, and female mating rate (Methods). As predicted by theory, coevolutionary change of the absolute level of arms escalation was in no case related to the outcome of sexually antagonistic interactions, whereas changes in the relative armament of the two sexes was associated strongly with evolutionary alterations in the outcome of sexually antagonistic interactions (see Table 1).

Collectively, these results show that the average coevolutionary trajectory that we have identified describes an evolutionary path along which levels of arms in the two sexes is to a great extent balanced. Had coevolution proceeded precisely along this path, the effects of the arms race on mating interactions would have been obscured. But species evolve off this path to points where one sex gains relative advantage over the other, and it is these deviations that provide insight into the consequences of this arms race. Species thus differ considerably both in the escalation of arms and in relative armament of the sexes. Theory suggests that such variation can be caused by several different factors, such as differences in armament costs, environmental variation, or intraspecific variation in armament levels^{13–15}.

A closer inspection of the analyses (see Table 1) yielded four important insights. First, significant effects of relative, but not absolute, levels of armament were found consistently across all measures of the outcome of antagonistic interactions, which suggests that the sexual arms race propels evolution of the general mating system in this group of insects³⁰. Second, our results were not affected by whether we used phylogenetically independent contrast or species-level data (Supplementary Information), which indicates that the degree of phylogenetic inertia is very low in our data²³. The phylogenetic autocorrelations^{16,24}, which range from -1 to $+1$, observed at the species level were also generally very low (scaled Moran's I (ref. 24) for male persistence, -0.544 ; for female resistance, 0.278 ; range for the four behavioural variables, -0.03 to 0.393 ; $P > 0.110$ in all cases), confirming that phenotypic similarity across species is at most only weakly related to shared ancestry. Such a weak relationship between phylogenetic similarity and phenotypic similarity implies that the rate of character evolution is relatively high and non-directional^{18,23,24}. Our results thus illustrate the dynamic nature of sexually antagonistic coevolution^{13–15}, with periods of escalation and de-escalation²⁰.

Third, evolutionary changes in the relative armament of the sexes accounted for a large proportion of the variation in evolution of interactions between males and females. Correlation analyses, using phylogenetically independent contrasts, showed that relative armament of the sexes alone (Table 1, PC2) explained 48, 41, 40 and 26% of the variance in our four behavioural variables, respectively. Finally, the sign of the relationship between relative armament of the sexes and the behavioural variables consistently corresponds to those predicted by *a priori* hypotheses. As females evolve a relative advantage in arms over males, they are able to dislodge harassing males more rapidly and more successfully in pre-mating struggles. Consequently, females evolve to mate less frequently and spend less time in superfluous matings.

The results of our analyses provide empirical evidence for a coevolutionary arms race between the sexes and for its consequences for the antagonistic interactions associated with reproduction¹³. Our comparative analytical strategy allowed us to confirm that sexually antagonistic coevolution may be hidden by continuous

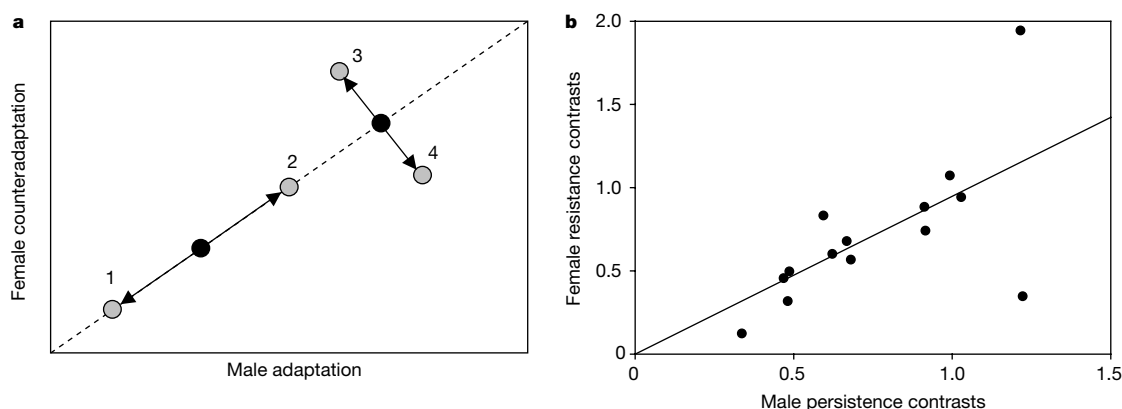


Figure 1 Coevolution of arms in males and females. **a**, Sexually antagonistic coevolution is expected to be hidden, because adaptations in one sex should be balanced by counteradaptation in the other. Mutual and matched de-escalation (1) or escalation (2) of armament in the two sexes should not therefore affect the outcome of antagonistic interactions. But when this arms race is not balanced perfectly, coevolution will lead to either females (3) or males (4) gaining a relative advantage in conflict. **b**, Water striders engage in pre-mating struggles. Species can be ordinated in a 2D space constructed by sexually antagonistic morphological adaptations in males and females that mediate this conflict²⁰, as identified by a two-block partial least-squares analysis²⁹ of male and female body shape, in which increases in magnitude represent increases in armament levels in both sexes. Shown are phylogenetically independent contrasts^{16,17} of such male

persistence and female resistance (solid line represents linear regression forced through the origin; test of $H_0: \beta = 0$, $t = 8.55$, $P < 0.001$; $r = 0.92$). Male persistence represents a multivariate measure of male expression of morphologies associated with male ability to secure a firm grip of females during pre-mating struggles. Female resistance represents a multivariate measure of female expression of morphologies that aid in mate rejection during pre-mating struggles. These 'arms' include, for example, elongation of grasping genitalia in males and abdominal spines in females that function to hinder grasping by males. The correlated evolution between male persistence and female resistance involves different traits in the two sexes, and is not caused by a common correlated evolution with body size (that is, size allometry²⁰).

adaptation and counteradaptation^{1–3}. It also shows how the evolutionary consequences of sexual conflict can be nevertheless determined, by analysing the effects of small imbalances in male and female adaptations that may occur during an arms race. Finally, our study illustrates the power of combining quantifications of adaptation and counteradaptation in both sexes with independent measures of the resulting interactions across populations or species. This comparative strategy will help us to understand the elusive but potent process of sexually antagonistic coevolution. □

Methods

We collected 15 congeneric water strider species (*Gerris* spp.) from various sites in Europe and North America during their peak reproductive season, and transported them to the laboratory for the experiments described below. These individuals were also subjected to morphological analyses. A list of species and detailed descriptions of the collection sites and morphometric methods is provided elsewhere²⁰.

Behavioural experiments

Interspecific behavioural variation in antagonistic mating interactions were quantified in two sets of replicated laboratory experiments conducted separately on each of the 15 species. Water striders were kept in aerated pools (water depth 3–6 cm) provided with Styrofoam strips at 20 ± 1 °C, and were fed frozen insects (fruit flies and crickets). We made every possible effort to standardize the biotic and abiotic protocol under which these large experiments were carried out. In the first set, eight individuals were introduced into each of 6–10 replicate experimental pools (40 × 60 cm) per species at a sex ratio of 3:1. By continuously observing all interactions (for 1 h), we measured the average duration of pre-mating struggles that led to dislodgement of the male and the proportion of struggles in which females yielded in each replicate pool. The latter struggles led to mating and hence measure male struggle success.

In the second set of experiments, eight individually marked water striders were placed in each of 14–20 replicate experimental pools (40 × 60 cm) per species at an average sex ratio of 1:1 and were allowed to acclimatize to these conditions for a period of 24 h. We then measured female mating activity (the mean proportion of time spent mating) and female mating rate, by conducting spot checks at 10-min intervals in all replicates (total observation time per replicate: 9–25 h over two consecutive days).

All analyses presented were carried out using species averages across all replicates. Note that because both struggles and superfluous mating are costly for females^{25,26}, the evolution of reduced magnitudes of all four behavioural variables is clearly in the interest of females.

Statistical methods

We based all statistical inference described below on phylogenetically independent contrasts¹⁷ of all variables, which we computed using the CONTRAST module of PHYLIP²⁷ and a phylogenetic hypothesis that is based on both molecular (820 base pairs (bp) from the mitochondrial COI gene and 515 bp from the nuclear EF-1 α gene) and morphological (63 characters) data²⁸. These contrasts account for bias that may be introduced by shared ancestry among our species, and thus allow direct comparative analyses of correlated evolution^{16,17,23}. No branch lengths have been estimated for our phylogenetic hypothesis, so equal branch lengths were assumed^{16,17}. In no case was the absolute magnitude of the contrasts correlated (Pearson product–moment correlations) significantly with the estimated standard deviation (ref. 17; $0.074 \leq P \leq 0.822$ in all cases). To test whether the outcome of antagonistic interactions between the sexes coevolves with the evolution of morphological arms in the two sexes, we analysed our data using two different approaches. First, we fitted male persistence and female resistance simultaneously as independent variables to each of the behavioural variables in a series of multiple regression models. Because each independent variable in a multiple regression model estimates effects that are independent of other independent variables included in the model, these models estimate the effects of both male persistence and female resistance, given that the other variable is held constant. In these models, we therefore predict that an increase in male persistence (that is, $\beta > 0$) and a decrease in female resistance (that is, $\beta < 0$) will both be associated with increases in our behavioural variables, because such changes both result in a relative male advantage in arms.

Second, we tested for the independent effects of (1) absolute level of arms, or the position along the average coevolutionary trajectory, and (2) relative armament of the sexes, or the degree to which the level of arms is mismatched in the two sexes (Fig. 1a), using the following procedure. To generate statistically independent measures of these two parameters, we first subjected the two correlated variables, male persistence and female resistance (see Fig. 1b), to a principal component analysis, which was based on the correlation matrix to ensure that variance in male persistence and female resistance was given equal weight, and which yielded two principal components (PCs). The first (PC1) was highly correlated with both male persistence ($r = 0.98$) and female resistance ($r = 0.98$), and thus provides an integrative measure of the absolute level of arms in males and females. The second (PC2) is by definition orthogonal to (that is, uncorrelated with) the first, and hence directly measures the degree of mismatch, or imbalance, in arms level between the two sexes: PC2 was negatively correlated with male persistence ($r = -0.20$) and positively with female resistance ($r = 0.20$).

We carried out a series of multiple regression analyses using PC1 and PC2 as independent variables, and behaviour as the dependent variable. Because a positive score

on PC2 corresponds to a relative female advantage in arms (Fig. 1a, 3), we predict that there is a negative coevolutionary relationship between PC2 and our behavioural variables ($\beta < 0$). For comparative purposes only, we also carried out these analyses with species-level rather than contrast data (Supplementary Information).

Received 22 August; accepted 29 November 2001.

- Chapman, T. & Partridge, L. Sexual conflict as fuel for evolution. *Nature* **381**, 189–190 (1996).
- Rice, W. R. Sexually antagonistic male adaptation triggered by experimental arrest of female evolution. *Nature* **381**, 232–234 (1996).
- Rice, W. R. Dangerous liaisons. *Proc. Natl Acad. Sci. USA* **97**, 12953–12955 (2000).
- Partridge, L. & Hurst, L. D. Sex and conflict. *Science* **281**, 2003–2008 (1998).
- Gavrilets, S., Arnqvist, G. & Friberg, U. The evolution of female mate choice by sexual conflict. *Proc. R. Soc. Lond. B* **268**, 531–539 (2001).
- Civetta, A. & Clark, A. G. Correlated effects of sperm competition and postmating female mortality. *Proc. Natl Acad. Sci. USA* **97**, 13162–13165 (2000).
- Gavrilets, S. Rapid evolution of reproductive barriers driven by sexual conflict. *Nature* **403**, 886–889 (2000).
- Swanson, W. J., Yang, Z., Wolfner, M. F. & Aquadro, C. F. Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. *Proc. Natl Acad. Sci. USA* **98**, 2509–2514 (2001).
- Arnqvist, G. Comparative evidence for the evolution of genitalia by sexual selection. *Nature* **393**, 784–786 (1998).
- Rice, W. R. in *Endless Forms—Species and Speciation* (eds Howard, D. J. & Berlocher, S. H.) 261–270 (Oxford Univ. Press, Oxford, 1998).
- Parker, G. A. & Partridge, L. Sexual conflict and speciation. *Phil. Trans. R. Soc. Lond. B* **353**, 261–274 (1998).
- Arnqvist, G., Edvardsson, M., Friberg, U. & Nilsson, T. Sexual conflict promotes speciation in insects. *Proc. Natl Acad. Sci. USA* **97**, 10460–10464 (2000).
- Parker, G. A. in *Sexual Selection and Reproductive Competition in Insects* (eds Blum, M. S. & Blum, N. A.) 123–163 (Academic, New York, 1979).
- Parker, G. A. Arms races in evolution—an ESS to the opponent-independent cost game. *J. Theor. Biol.* **101**, 619–648 (1983).
- Härdling, R. Arms races, conflict costs and evolutionary dynamics. *J. Theor. Biol.* **196**, 163–167 (1999).
- Harvey, P. H. & Pagel, M. D. *The Comparative Method in Evolutionary Biology* (Oxford Univ. Press, Oxford, 1991).
- Garland, T., Harvey, P. H. & Ives, A. R. Procedures for the analysis of comparative data using phylogenetically independent contrasts. *Syst. Biol.* **41**, 18–32 (1992).
- Rowe, L., Arnqvist, G., Sih, A. & Krupa, J. J. Sexual conflict and the evolutionary ecology of mating patterns: water striders as a model system. *Trends Ecol. Evol.* **9**, 289–293 (1994).
- Arnqvist, G. in *The Evolution of Mating Systems in Insects and Arachnids* (eds Choe, J. C. & Crespi, B. J.) 146–163 (Cambridge Univ. Press, Cambridge, 1997).
- Arnqvist, G. & Rowe, L. Correlated evolution of male and female morphologies in water striders. *Evolution* (in the press).
- Arnqvist, G. Sexual selection in water strider: the function, nature of selection and heritability of a male grasping apparatus. *Oikos* **56**, 344–350 (1989).
- Arnqvist, G. & Rowe, L. Sexual conflict and arms races between the sexes: a morphological adaptation for control of mating in a female insect. *Proc. R. Soc. Lond. B* **261**, 123–127 (1995).
- Losos, J. B. Uncertainty in the reconstruction of ancestral character states and limitations on the use of phylogenetic comparative methods. *Anim. Behav.* **58**, 1319–1324 (1999).
- Gittleman, J. L. & Kot, M. Adaptation—statistics and a null model for estimating phylogenetic effects. *Syst. Zool.* **39**, 227–241 (1990).
- Rowe, L. 1994. The costs of mating and mate choice in water striders. *Anim. Behav.* **48**, 1049–1056 (1994).
- Watson, P. J., Arnqvist, G. & Stallman, R. R. Sexual conflict and the energetic costs of mating and mate choice in water striders. *Am. Nat.* **151**, 46–58 (1998).
- Felsenstein, J. The Phylogeny Inference Package [online]. (<http://evolution.genetics.washington.edu/phylip.html>) (2001).
- Damgaard, J. & Sperling, F. A. H. Phylogeny of the water strider genus *Gerris* Fabricius (Heteroptera: Gerridae) based on COI mtDNA, EF-1 α nuclear DNA and morphology. *Syst. Entomol.* **26**, 241–254 (2001).
- Rohlf, F. J. & Corti, M. Use of two-block partial least-squares to study covariation in shape. *Syst. Biol.* **49**, 740–753 (2000).
- Rowe, L. & Arnqvist, G. Sexually antagonistic coevolution in a mating system: combining comparative and experimental approaches to address evolutionary process. *Evolution* (in the press).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com>).

Acknowledgements

This study was supported by the Swedish Natural Science Research Council, the Natural Sciences and Engineering Research Council of Canada, the Swedish Foundation for International Cooperation in Research and Higher Education, the Knut and Alice Wallenberg Foundation and the Magnus Bergvalls Stiftelse. We thank N. M. Andersen and J. Damgaard for phylogenetic information; J. Felsenstein and F. J. Rohlf for developing software; and T. Day, U. Friberg, D. Schluter and T. Tregenza for comments.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.A. (e-mail: Goran.Arnqvist@ebc.uu.se).

Context-enabled learning in the human visual system

Yael Adini, Dov Sagi & Misha Tsodyks

Department of Neurobiology, Brain Research, The Weizmann Institute of Science, Rehovot 76100, Israel

Training was found to improve the performance of humans on a variety of visual perceptual tasks^{1,2}. However, the ability to detect small changes in the contrast of simple visual stimuli could not be improved by repetition³. Here we show that the performance of this basic task could be modified after the discrimination of the stimulus contrast was practised in the presence of similar laterally placed stimuli, suggesting a change in the local neuronal circuit involved in the task. On the basis of a combination of hebbian and anti-hebbian synaptic learning rules compatible with our results, we propose a mechanism of plasticity in the visual cortex that is enabled by a change in the context.

Practice is known to improve the ability of humans to detect small changes in a variety of visual attributes such as position, orientation, motion, depth, spatial phase, hyper-acuity and texture segmentation^{1,2}. For these kinds of tasks, repetitions have resulted in a gradual improvement in the performance (perceptual learning), until a performance plateau (saturation) is reached at a new level. The specificity of these learning effects to the spatial position, to the orientation of the stimuli and to the eye that receives the stimuli has served as evidence of long-lasting plasticity in the primary visual cortex of human adults^{4–6}. Accumulating evidence^{7,8} shows changes in the primary visual cortex of the monkey that are correlated with learning effects, supporting the psychophysical results.

Assuming that adults retain plasticity at low levels of visual processing, we would expect that practising the ability to notice small changes in the stimulus contrast (contrast-discrimination task, or CD) would also lead to an improvement. However, CD is an example of a task where practice does not 'make perfect'. In refs 3 and 9, observers repeated CD experiments over a long period of time (over 40 sessions), without any improvement. In fact, the only observable effect of the practice was a slight deterioration in the performance, indicating that this basic task¹⁰ might have reached its optimal performance either during normal development or within the first 500 trials that composed a testing session (that is, during 'fast-learning processes'^{6,11}).

The saturation of perceptual learning for most of the tasks, as well as the absence of learning for the CD task, indicates that not all the activations of the primary visual cortex result in long-term modifications. Moreover, it appears that any particular pattern of stimulation ceases to become effective in evoking modifications when it becomes 'familiar' to the visual system. Thus, the saturated performance following repetitions may not necessarily reflect the best possible performance of a task. To explore this possibility, we searched for practice conditions that would improve the seemingly saturated performance of the CD task. We introduced a set of CD practice sessions in which the target stimulus was surrounded by chains of similar stimuli of fixed high contrast and varying length¹² (Fig. 1b). We then compared the CD performance for isolated target stimuli, measured before and after the practice sessions.

Six adult observers, with typical contrast-discrimination abilities¹³ participated in the experiments. After measuring their performance in the contrast discrimination of an isolated foveal Gabor signal^{14,15}, the observers practised the same task in the presence of chains of high-contrast lateral Gabor signal flankers (Fig. 1). The practice was organized in subsequent cycles of three daily sessions each, with 500–1,000 trials a day. During each

experimental sub-session, the contrast of the target stimulus was varied from 0 to 50%, while the length of the chain of flankers was held constant; this length was increased (2–10) from one sub-session to another. After 2–3 practice cycles, all observers had improved their discrimination performance. Most importantly, the CD thresholds for isolated target stimuli, tested after each practice cycle, were significantly reduced (Fig. 2a). Specifically, there was a marked decrease in the just-noticeable difference in contrast (the threshold) with reference contrasts above 10%. On average ($N = 6$ observers), the CD threshold for a reference contrast of 50% decreased to about half of its initial value (by $0.28 \pm \text{s.d.} = 0.03$ logunits; s.d., standard deviation). This decrease was highly significant ($P < 0.0002$, $N = 6$). A significant decrease ($P < 0.002$, $N = 6$) was also found in the CD threshold for a reference contrast of 25%, where the threshold decreased to 0.58 of its initial value ($0.24 \pm \text{s.d.} = 0.04$ logunits).

The effect of the lateral flankers used in our learning procedure could be merely to increase the overall excitation at the target location. To control for this possibility we used the following procedure: two observers practised the discrimination of a partial range of high contrasts for a number of trials similar to that of the

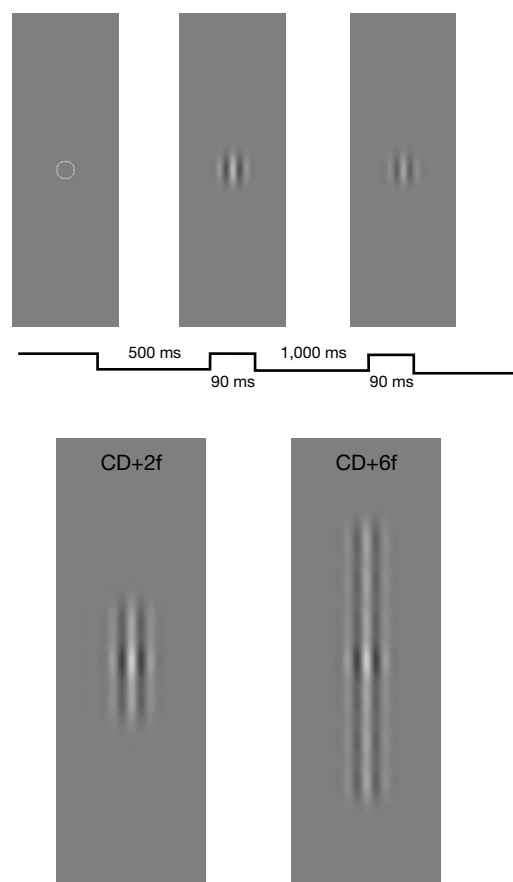


Figure 1 Experimental conditions. **a**, In the contrast-discrimination (CD) task we measured the just-noticeable difference in contrast needed to discriminate the two Gabor signals. The observers report which of the two stimuli appears to have a higher contrast. A temporal two-alternative forced-choice procedure was used, as described by the time sequence in the figure, and in more detail in the Methods. **b**, Examples of stimuli used during the 'practice with flankers' sessions, in which the thresholds for contrast-discrimination for the central Gabor signal were measured in the presence of chains of collinear flankers. Here we show a chain of two flankers (2f), and a chain of six flankers (6f).

first experiment. This learning procedure did not result in any statistically significant change in the performance of the CD task (Fig. 2b), which implies that the lateral input from the proximal flankers is not equivalent to direct input to the target location^{16,17}.

The time course of the learning process can be evaluated by monitoring the changes in the CD performance after subsequent practice cycles. As illustrated in Fig. 3, most of the learning effect took place after one cycle (three sessions taken on different days) of practice in the presence of a chain of lateral GS flankers. We note that a similar time course of learning was found with other perceptual tasks where the new level of performance was reached after 4–6 sessions taken on different days¹¹. The learning effect was found to last for at least eight months (Fig. 3), indicating long-lasting modifications in the human visual cortex^{4,11}.

To account for the learning effect that was found here, we have assumed that the CD performance involves the activation of an interconnected local neural network^{18,19}, with synaptic connections that are modified in an activity-dependent manner. Because the performances were found to be stable after repetitions with different contrasts, but improved after practising the task with a new context, we looked for a plasticity mechanism that is activated by a change of context. Such a mechanism can be sub-served by the properties of long-term synaptic modification analysed recently²⁰. In studying the implications of spike-time-dependent synaptic plasticity^{21–23}, the following approximate expression for changing the probability of neurotransmitter release (P_r ; equivalent to the 'strength') for a synaptic connection was formulated

$$\frac{dP_r}{dt} = r_1 P_r f_{pre} f_{post}^2 - r_2 (P_r f_{pre})^2 f_{post} \quad (1)$$

This function of the pre- and post-synaptic instantaneous firing frequencies ($f_{pre}(t)$ and $f_{post}(t)$) combines nonlinear hebbian and anti-hebbian terms²⁴ that determine the up- and downregulation of the release probability, and thus the effectiveness of the synaptic connection. (Here r_1 and r_2 denote the corresponding rate constants). By rewriting equation (1) in a slightly different way, we obtain

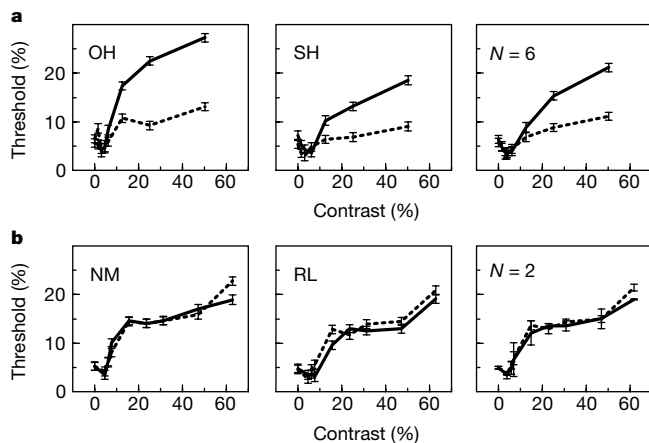


Figure 2 Changes in the contrast-discrimination curves. Data were obtained after **a**, practising contrast-discrimination in the presence of chains of flankers (a new context) and **b**, practising discrimination of high-contrast stimuli (the same context). For each condition the graph shows the data for two individual observers, and the average across N observers. Each datum point in the observers' data is the average of 2–4 threshold estimates taken on different days. The solid curves represent the data before the practice. The dotted curves represent the after-practice data.

$$\frac{dP_r}{dt} = r_1 P_r f_{pre} f_{post} (f_{post} - P_r \epsilon f_{pre}) \quad (2)$$

where ϵ is defined as r_2/r_1 . According to this equation, a steady state is reached when

$$P_r = \frac{f_{post}}{\epsilon f_{pre}} \quad (3)$$

Thereafter, the release probability will remain steady as long as the ratio between the post- and pre-synaptic rate remains constant. However, when the rates deviate from this relation, the release probability begins to change until a new equilibrium point is reached (Fig. 4a). The release probability of the synapse can therefore serve as a memory for the expected relation between the pre- and post-synaptic firing rates. We next show how such a mechanism can explain the stability of human CD performance when practised with stimuli of different contrasts, and the change in the performance when a new stimulus configuration is introduced.

We model the CD task as mediated by a local cortical column consisting of two interconnected subpopulations of excitatory and inhibitory neurons^{12,25} (Fig. 4b and see the Methods). The activity of the excitatory (E) and the inhibitory (I) subpopulations is determined by the external inputs, which are increasing functions of the stimuli contrast, and the recurrent interactions in the local network. When the contrast of the visual stimulus is increased, the resulting activity (E) also increases, enabling the discrimination between the contrasts. The CD threshold is controlled by the steepness of the relation between the activity (E) and the contrast. Our psychophysical results could be explained in this model if synaptic connections are modified according to the principles presented above. Assume that the network receives stimulus-driven thalamic inputs (e and i) that are divided between the excitatory and inhibitory subpopulations in a certain fixed proportion that is, $i = ke$ (ref. 12). We can show that in the linear approximation, the ratio between the average firing rates E and I is contrast independent (see Methods). Thus, according to the above analysis of synaptic

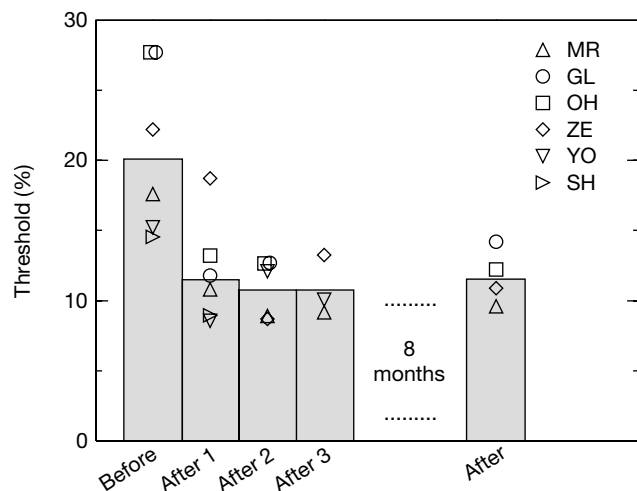


Figure 3 The time course of the context-induced learning. The figure shows the development with practice of the changes in the contrast-discrimination thresholds with a reference contrast of 50%. Thresholds measured before and after consecutive cycles of practice are shown for six observers, with their average represented by the histogram bars. Observers practised the collinear configurations, except observer MR, who practised the parallel configurations, obtaining similar results. Each 'after- N ' measurement was taken on the day after the N th practice cycle. Four observers were tested after about 8 months to check the retention of their learning effect. The average of their average thresholds is represented with the rightmost bar ('after'). The symbols represent the thresholds obtained for each of the six observers.

plasticity, the strength of the connections, when reaching equilibrium levels, will not change further when the same local visual stimulus is repeatedly shown at various contrasts (Fig. 4a). If, however, a new stimulus configuration is presented, such as the one with the laterally placed flankers, synaptic modification will be rekindled. This is because the lateral input from the flankers may be distributed differently between the excitatory and inhibitory subpopulations and this will change the E/I ratio. According to equation (3), the equilibrium values for the corresponding connections (J_{ei} and J_{ie}) will change after the presentation of the flankers. To make the analysis simpler we assume that only one type of connection undergoes modification. In particular, if the lateral input is more biased towards inhibition ($\Delta i = k_1 \Delta e$; $k_1 > k$), which is compatible with our previous results¹² and with physiological findings²⁶, then the overall ratio between the inputs to inhibitory and excitatory subpopulations (k) will be larger

$$k \rightarrow \frac{i + \Delta i}{e + \Delta e} = k + (k_1 - k) \frac{\Delta e}{e + \Delta e} \quad (4)$$

Calculations show that in this case the new stationary value of J_{ei} (or J_{ie}) will be lower (Methods), resulting in a steeper dependency of the activity of the excitatory subpopulation E , on the contrast. The CD threshold is therefore lowered, as demonstrated by the model simulations (Fig. 4c).

Our psychophysical findings show that the seemingly stable contrast discrimination (CD) performance could be modified after practising the task in the presence of similar flankers. The system is thus capable of long-term adjustments in its local gain, according to changes in the relevant environment. This result demonstrates that the visual system possesses a mechanism for

context-induced learning. This mechanism allows the visual system to distinguish between stimulus changes that are irrelevant for object recognition and learning (for example, stimulus contrast), and changes affecting object shape (for example, stimulus configuration). Inspired by recent physiological results^{21–23}, we suggest that this type of learning is mediated by a synaptic modification rule that combines hebbian and anti-hebbian types of plasticity and is activated by changes in the relative proportions of the excitatory and inhibitory activities in the local visual circuits. Our data suggest that the specific context that was practised here leads to decreased inhibitory influences. We expect that other practice contexts will be found which will lead to an opposite change in the relative excitatory-inhibitory activities, and thus to higher inhibitory influences.

Our learning procedure resulted in an increased sensitivity to changes in contrast, and so it may be used to improve the contrast discrimination ability in humans. A challenging study would be to apply variations of our learning technique to improve the sensitivity of other sensory attributes. □

Methods

Apparatus

Stimuli were displayed as grey-level modulation on a Philips colour monitor, using a personal computer with an Intel Pentium II processor. For full details see ref. 16.

Stimuli

The stimuli consisted of one target signal (at the fixation point), one reference signal (at the target location) and 0, 2, 4, 6, 8 or 10 lateral masks (flankers). The spatial luminance distribution of each of the target, reference and mask signals was described by a Gabor function (a cosine grating multiplied by a gaussian envelope²⁷, with a vertical orientation, and $\sigma = \lambda = 0.15^\circ$). The reference-signal amplitude, C_r , was changed from 0 up to 63% during the experimental conditions. The amplitude of the Gabor signal was taken to represent its contrast. The target and masks were aligned with 2λ (0.3°) spacing.

Experimental procedure

A temporal two-alternative forced choice procedure was used (Fig. 1a). A staircase method³ was used to determine the contrast threshold at a level of 79% correct. For full details see ref. 16.

Psychophysical experiments

Observers practiced the CD task using either the 'practice with flankers' or the 'practice without flankers' learning procedures (see below). The results were described by a threshold versus contrast (TVC) function. After measuring the initial TVC curve for each observer (2–4 repetitions), observers practised CD tasks using a specific learning procedure. Practice was taken in cycles. Each practice cycle was taken on three successive days. Observers participated in 2–3 practice cycles. One TVC curve was taken for each observer the day after each practice cycle. We used the following learning procedures.

In experiment 1 (practice with flankers), contrast-discrimination thresholds were measured for a foveal Gabor signal flanked on each side by either 1, 2, 3, 4 or 5 high-contrast (30%) Gabor signal masks. The reference contrast varied from zero to 50% ($C_r = 0, 1.5, 3, 6, 12.5, 25$ and 50%) during each experimental sub-session, whereas the flankers' contrast ($C_m = 30\%$) and the number of flankers were fixed. Observers usually participated in two 20-min experimental sub-sessions each time they came to the laboratory. They had a 20-min break between the successive sub-sessions. One cycle of practice consisted of five experimental sub-sessions (each with a different number of flankers) and was taken on three successive days. Six adult observers (17–24 years old) participated in this experiment. Five observers practised with collinear flankers and one observer practised with parallel flankers. An observer (S.H.) practised the discrimination of just two reference-contrasts (0 and 30%) in the presence of either two or four high-contrast lateral flankers.

In experiment 2 (practice without flankers), two 17-year-old observers participated in this learning procedure. During each session, the reference contrast was varied over a limited range of supra-threshold contrasts ($C_r = 0, 30\%, 47\%$ and 63%). A session consisted of 10 blocks of CD experiments, followed by two blocks of CD experiments ($C_r = 0, 30\%$) that were performed about 20 min after the end of the daily session. The last block of the daily session used a high-contrast reference stimulus (30%). One cycle of practice contained three sessions taken on three successive days.

Network model of contrast discrimination

The network model is illustrated in Fig. 4b. Assuming the threshold-linear gain functions for both subpopulations, for given input ($e, i = ke$), the activities converge to a steady-state solution of the following equations^{12,28}

$$\begin{aligned} E &= e - J_{ei}I + J_{ee}E \\ I &= ke - J_{ii}I + J_{ie}E \end{aligned} \quad (5)$$

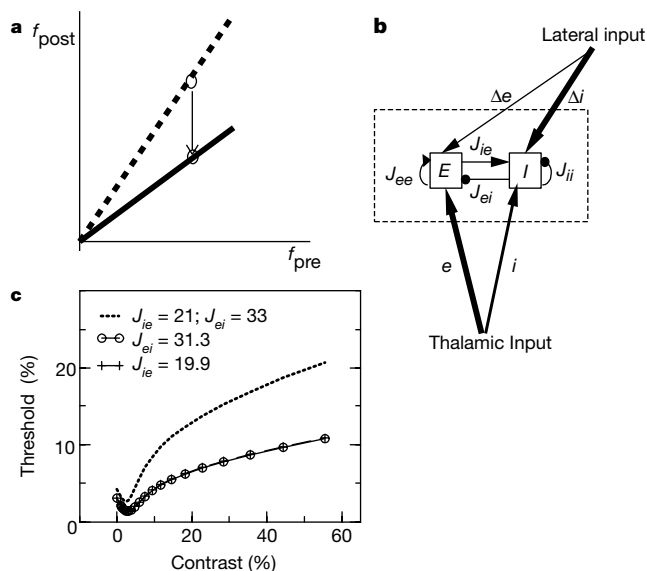


Figure 4 Putative synaptic mechanism and model simulations. **a**, Linear relationship between pre- and post-synaptic frequencies for which synaptic modification does not occur. The slope of the line depends on the synaptic strength (probability of release). Deviating from the linear relationship initiates a modification of the release probability, until it reaches a new stable value. **b**, A schematic representation of a cortical hyper-column¹², consisting of two interconnected subpopulations: excitatory (E) and inhibitory (I). The symbols e and i denote the strength of the thalamic sensory inputs to these two subpopulations; Δe and Δi are lateral inputs. The width of the corresponding lines illustrates the assumed differences in the relative distribution of inputs to excitatory and inhibitory subpopulations. **c**, Performance of the model on the CD task before (dotted) and after (solid) modifications in either of the connections J_{ei} or J_{ee} (which led to practically the same effect). See Methods for details; compare with Fig. 2a.

where J_{ei} denotes the average strength of I to E connections (analogously for the other connections), and is an increasing function of the stimulus contrast (see below). We have assumed that the steady-state solution of equation (5) is stable and both E and I are positive for a positive c , which imposes certain conditions on the values of the connection strengths (see for example, ref. 29). In particular, if recurrent excitation exceeds the destabilizing level ($J_{ee} > 1$), it has to be offset by strong enough inhibition. Under these conditions, the CD is determined by the dependency of the activity E on the contrast

$$E = e \frac{1 + J_{ii} - kJ_{ei}}{J_{ei}J_{ie} - (J_{ii} + 1)(J_{ee} - 1)} \quad (6)$$

Simple algebra shows that this dependency becomes steeper when the strength of either the inhibitory connection J_{ei} or the excitatory connection J_{ie} decreases, thus leading to better CD performance. We assume that one of these connection types undergoes modification according to equation (1). For example, for the connection between the subpopulations I and E we write explicitly $f_{pre} = I, f_{post} = E$ and $J_{ei} = P_i W_{ei}$, where W is a constant factor. It follows from equation (5) that the ratio between the rates of the subpopulations is independent of the input c (that is, on the stimulus contrast)

$$\frac{E}{I} = \frac{1 + J_{ii} - kJ_{ei}}{J_{ie} - k(J_{ee} - 1)} \quad (7)$$

Thus, according to equations (3) and (7), if the strength of the connections is at the equilibrium values, activation of the network at different contrasts does not lead to synaptic modifications. However, the modification will occur if the value of k is changed. Calculations show that when k increases owing to laterally placed flankers (equation (4)), a new equilibrium is reached at which the strength of the connection that undergoes the modification (either J_{ei} or J_{ie}) is weaker. In order for the equilibrium to be stable, the sign of the rate constants (r_1, r_2) in equation (1) have to be positive for J_{ei} and negative for J_{ie} ; that is, these different types of connections should obey modification rules with exchanged powers of the corresponding hebbian and anti-hebbian terms.

To simulate our psychophysical data, we chose the widely used^{18,19,30} Naka–Rushton function³¹ to describe the thalamic input

$$e = \frac{C^p}{C^q + A^q} \quad (8)$$

where C represents the stimulus contrast. In the simulations (Fig. 4c) we used $p = 3.5$, $q = p - 0.5$, and $A = 3.5$. The thalamic input to the I node was $i = 0.3e$. The simulations for the 'before' practice curve were obtained using equation (5) with $J_{ee} = J_{ie} = 21$, $J_{ii} = 30$ and $J_{ei} = 33$. The 'after' simulations were obtained using either $J_{ei} = 31.3$, or $J_{ie} = 19.9$. All the other parameters were the same as in the 'before' simulations. We have assumed that any two contrasts, C_1 and C_2 , can be discriminated if $E(C_1) - E(C_2) \geq 0.4$ (the criterion for discrimination was fixed according to the value of the CD at zero contrast).

Received 29 October; accepted 29 November 2001.

- Sagi, D. & Tanne, D. Perceptual learning: learning to see. *Curr. Opin. Neurobiol.* **4**, 155–159 (1994).
- Karni, A. The acquisition of perceptual and motor skills: a memory system in the adult human cortex. *Brain Res. Cogn. Brain Res.* **5**, 39–48 (1996).
- Dorais, A. & Sagi, D. Contrast masking effects change with practice. *Vision Res.* **37**, 1725–1733 (1997).
- Karni, A. & Sagi, D. Where practice makes perfect in texture discrimination-evidence from primary visual cortex plasticity. *Proc. Natl Acad. Sci. USA* **88**, 4966–4970 (1991).
- Polat, U. & Sagi, D. Spatial interactions in human vision: from near to far via experience-dependent cascades of connections. *Proc. Natl Acad. Sci. USA* **91**, 1206–1209 (1994).
- Fahle, M., Edelman, S. & Poggio, T. Fast perceptual learning in hyperacuity. *Vision Res.* **35**, 3003–3013 (1995).
- Schoups, A., Vogels, R. & Orban, N. Q. G. Practicing orientation identification improves orientation coding in V1 neurons. *Nature* **412**, 549–553 (2001).
- Crist, R. E., Li, W. & Gilbert, C. D. Learning to see: experience and attention in primary visual cortex. *Nature Neurosci.* **4**, 519–525 (2001).
- Zenger, B. & Sagi, D. In *Textbook on 'Perceptual Learning'* (eds Fahle, M. & Poggio, T.) Ch. 10, 177–196 (MIT Press, Boston, 2002).
- Boynton, G. M., Demb, J. B., Glover, G. H. & Heeger, D. J. Neuronal basis of contrast discrimination. *Vision Res.* **39**, 257–269 (1999).
- Karni, A. & Sagi, D. The time course of learning a visual skill. *Nature* **365**, 250–252 (1993).
- Adini, Y., Sagi, D. & Tsodyks, M. Excitatory-inhibitory network in the visual cortex: psychophysical evidence. *Proc. Natl Acad. Sci. USA* **94**, 10426–10431 (1997).
- Legge, G. E. A power law for contrast discrimination. *Vision Res.* **21**, 457–467 (1981).
- Marcelja, S. Mathematical description of the responses of simple cortical cells. *J. Opt. Soc. Am.* **70**, 1297–1300 (1980).
- Pollen, D. A. & Romer, S. F. Visual cortical neurons as localized spatial frequency filters. *IRRR Trans. Syst. Man Cybern. SMC-13*, 907–916 (1983).
- Adini, Y. & Sagi, D. Recurrent networks in human visual cortex: psychophysical evidence. *J. Opt. Soc. Am. A* **18**, 2228–2236 (2001).
- Zenger-Landolt, B. & Koch, C. Flanker effects in peripheral contrast discrimination—psychophysics and modeling. *Vision Res.* **41**, 3663–3675 (2001).
- Wilson, H. R. & Humanski, R. Spatial frequency adaptation and contrast gain control. *Vision Res.* **33**, 1133–1149 (1993).
- Foley, J. M. Human luminance pattern-vision mechanisms: Masking experiments require a new model. *J. Opt. Soc. Am. A* **11**, 1710–1719 (1994).
- Senn, W., Markram, H. & Tsodyks, M. An algorithm for modifying neurotransmitter release probability based on pre- and postsynaptic spike timing. *Neur. Computat.* **13**, 35–67 (2001).
- Markram, H., Lübke, J., Frotscher, M. & Sakmann, B. Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science* **275**, 213–215 (1997).

- Bell, C. C., Han, V. Z., Sugawara, Y. & Grant, K. Synaptic plasticity in a cerebellum-like structure depends on temporal order. *Nature* **387**, 278–281 (1997).
- Bi, G. Q. & Poo, M. M. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *J. Neurosci.* **18**, 10464–10472 (1998).
- Abbott, L. F. & Nelson, S. B. Synaptic plasticity: taming the beast. *Nature Neurosci.* **3**, 1178–1183 (2000).
- Somers, D. C. *et al.* A local circuit approach to understanding integration of long-range inputs in primary visual cortex. *Cereb. Cortex* **8**, 204–217 (1998).
- Walker, G. A., Ohzawa, I. & Freeman, R. D. Suppression outside the classical cortical receptive field. *Vis. Neurosci.* **17**, 369–379 (2000).
- Gabor, D. Theory of communication. *J. Inst. Elect. Eng. (Lond.)* **93**, 429–457 (1946).
- Wilson, H. R. & Cowan, J. D. A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetik* **13**, 55–80 (1973).
- Tsodyks, M. V., Skaggs, W. E., Sejnowski, T. J. & McNaughton, B. L. Paradoxical effects of external modulation of inhibitory interneurons. *J. Neurosci.* **17**, 4382–4388 (1997).
- Albrecht, D. G. & Hamilton, B. Striate cortex of monkey and cat: contrast response function. *J. Neurophysiol.* **48**, 217–237 (1982).
- Naka, K. I. & Rushton, W. A. H. S-potentials from luminosity units in the retina of fish (cyprinidae). *J. Physiol. Lond.* **185**, 587–599 (1966).

Acknowledgements

We thank E. Ahissar, Y. Dudai and H. Markram for helpful comments on the manuscript. This research was supported by the Israeli Academy of Sciences, US-Israel Binational Foundation, the Office of Naval Research and the National Science Foundation.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.S. (e-mail: dov.sagi@weizmann.ac.il).

Excitatory glycine receptors containing the NR3 family of NMDA receptor subunits

Jon E. Chatterton^{*†}, Marc Awobuluyi^{*†}, Louis S. Premkumar[‡], Hiroto Takahashi^{*}, Maria Talantova^{*}, Yeonsook Shin[§], Jiankun Cui^{*}, Shichun Tu^{*}, Kevin A. Sevarino^{||}, Nobuki Nakanishi^{*}, Gang Tong^{*¶}, Stuart A. Lipton^{*} & Dongxian Zhang^{*}

^{*} Center for Neuroscience and Aging, The Burnham Institute, 10901 North Torrey Pines Road, La Jolla, California 92037, USA

[‡] Department of Pharmacology, Southern Illinois University School of Medicine, Springfield, Illinois 62702, USA

[§] Department of Molecular Pharmacology, Kanazawa University Graduate School of Medicine, Kanazawa, Ishikawa 920-8640, Japan

^{||} Department of Psychiatry, University of Connecticut Health Center, Farmington, Connecticut 06030, USA

[¶] Department of Neurosciences, University of California—San Diego, La Jolla, California 92093, USA

[†] These authors contributed equally to this work

The N-methyl-D-aspartate subtype of glutamate receptor (NMDAR) serves critical functions in physiological and pathological processes in the central nervous system, including neuronal development, plasticity and neurodegeneration^{1,2}. Conventional heteromeric NMDARs composed of NR1 and NR2A–D subunits^{3,4} require dual agonists, glutamate and glycine, for activation. They are also highly permeable to Ca^{2+} , and exhibit voltage-dependent inhibition by Mg^{2+} . Coexpression of NR3A with NR1 and NR2 subunits modulates NMDAR activity^{5–7}. Here we report the cloning and characterization of the final member of the NMDAR family, NR3B, which shares high sequence homology with NR3A. From *in situ* and immunocytochemical analyses, NR3B is expressed predominantly in motor neurons,

whereas NR3A is more widely distributed^{5,6}. Remarkably, when co-expressed in *Xenopus* oocytes, NR3A or NR3B co-assembles with NR1 to form excitatory glycine receptors that are unaffected by glutamate or NMDA, and inhibited by D-serine, a co-activator of conventional NMDARs. Moreover, NR1/NR3A or -3B receptors form relatively Ca²⁺-impermeable cation channels that are resistant to Mg²⁺, MK-801, memantine and competitive antagonists. In cerebrocortical neurons containing NR3 family members, glycine triggers a burst of firing, and membrane patches manifest glycine-responsive single channels that are suppressible by D-serine. By itself, glycine is normally thought of as an inhibitory neurotransmitter. In contrast, these NR1/NR3A or -3B 'NMDARs' constitute a type of excitatory glycine receptor.

We have cloned a complementary DNA that codes for the seventh member of the NMDAR family, NR3B, from rat brain. The polypeptide sequence contains 1,002 amino acids, corresponding to a calculated relative molecular mass of 109,000 (*M_r* 109K). Multiple sequence alignment of NR3B with other glutamate receptor subunits (see Supplementary Information) indicated that NR3B is closely related to a subunit we previously cloned, NR3A (47% identity), but shares less identity with NR1 and NR2 subfamilies (17–21%) and non-NMDAR families (13–16%). Analysis of human genome data revealed the human NR3B gene on chromosome 19. No further genomic sequences with homology to NMDAR subunits were identified by a BLAST search (<http://www.ncbi.nlm.nih.gov/BLAST>); therefore, NR3B is most likely the final member of the NMDAR family.

The major structural features of the NMDAR family are highly conserved in NR3B, including the S1 and S2 agonist binding domains and membrane-spanning regions (see Supplementary Information). A detailed comparison of residues involved in glycine binding to NR1 subunits^{8–10} and glutamate binding to NR2 subunits^{11,12} revealed that NR3B (or NR3A) could potentially

share ligand specificity with either or both of these classes of NMDAR subunit. However, within the M2 region, which lines the ion channel pore, NR3B differs significantly from NR1 and NR2 but is similar to NR3A. These features suggest that NR3A and -3B subunits may confer previously unknown ligand-binding and channel-gating properties to heteromeric NMDAR complexes.

We analysed the expression pattern of NR3B in adult rat brain and spinal cord by *in situ* hybridization. Radiolabelled NR3B-specific antisense probes detected NR3B messenger RNA predominantly in the ventral horn of the spinal cord and in the facial and trigeminal nuclei of the brainstem (Fig. 1a, b). Higher-resolution analyses using digoxigenin-labelled probes revealed that expression of NR3B mRNA in these regions is restricted primarily to motor neurons (Fig. 1c, d). Additionally, we detected NR3B mRNA in a subset of large cells in the vestibular nuclei and reticular formation of the brainstem (data not shown). Next, we examined the expression of NR3B protein. High sequence similarity between NR3B and NR3A within the region recognized by monoclonal antibody K35/40, which labels NR3A but not other previously identified NMDAR subunits¹³, suggested that this antibody should cross-react with NR3B. We confirmed this cross-reactivity by western blot¹³ (Fig. 1e), and examined expression of NR3B protein in NR3A-deficient mice⁷ by immunocytochemistry (Fig. 1f, g). NR3B protein was most prominent in motor neurons in the spinal cord and brainstem. In contrast, we had previously shown that NR3A was more widely distributed^{5,6}.

We examined functional properties of NR3B expressed in *Xenopus* oocytes using two-electrode voltage-clamp recording. Functional receptors, defined by the presence of ligand-evoked current, were assembled in oocytes co-injected with cRNAs coding for NR3B and NR1 but not with NR3B alone. We found that, unlike conventional (that is, NR1/NR2) NMDARs, glycine fully activated NR1/NR3B receptors without glutamate or NMDA (Fig. 2a). At –80 mV,

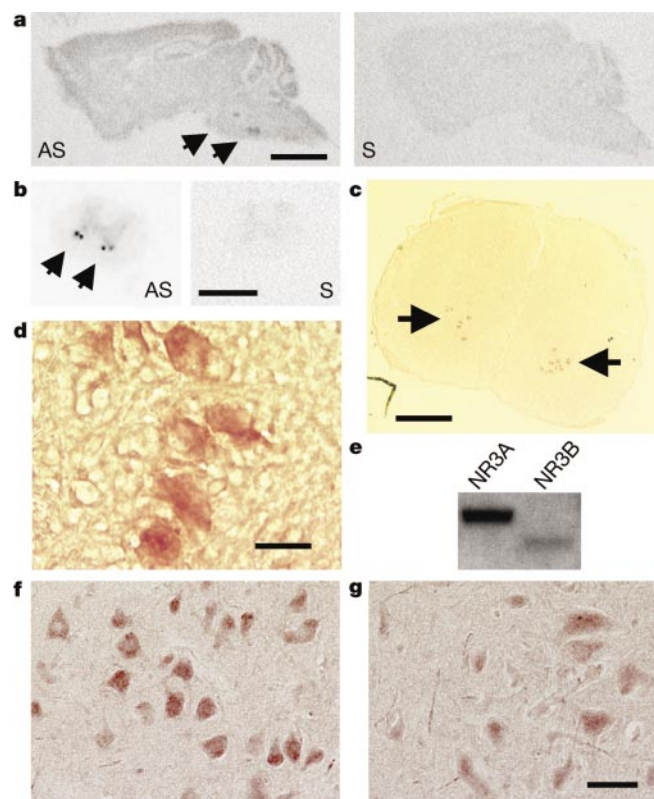


Figure 1 NR3B expression pattern by *in situ* hybridization and immunocytochemistry. **a**, **b**, Adult rat brain (**a**) or spinal cord (**b**) hybridized with ³³P-labelled antisense (AS) or sense (S) NR3B probes. **c**, **d**, Adult rat spinal cord hybridized with digoxigenin-labelled antisense NR3B probe. **e**, Western blot of NR3A- or NR3B-transfected HEK293T cell lysates probed

with antibody K35/40. **f**, **g**, Adult NR3A^{–/–} mouse facial nucleus (**f**) or spinal cord ventral horn (**g**) stained with antibody K35/40. Scale bars, 6 mm (**a**), 3 mm (**b**), 1 mm (**c**) or 50 μm (**d**, **f**, **g**). Arrows indicate specific labelling by the antisense probe.

application of 10 μ M glycine evoked inward currents up to several microamps. Importantly, L-glutamate and NMDA, agonists that activate conventional NMDARs through NR2 subunits, displayed little or no effect on glycine-activated currents of NR1/NR3B receptors, and did not evoke glycine-independent responses (Fig. 2b). In contrast, oocytes injected with NR3B alone or co-injected with NR3B and NR2A failed to express functional receptors. These results suggest that assembly of NR3B into a functional receptor complex minimally requires association with NR1. Previously, we and others^{5,6} were unable to detect functional recombinant NR1/NR3A receptors; however, this may have occurred because of the rapidly desensitizing nature of the glycine-evoked currents of NR1/NR3A receptors, as observed here (Fig. 2f). Consistent with previous reports^{3,4,14}, oocytes injected with NR1 alone expressed receptors that exhibited only very small currents (<10 nA) and required NMDA in addition to glycine for maximal activation (data not shown). Additionally, we and others have found that co-expression of NR1/NR2A/NR3B, similar to our previous findings with NR1/NR2A/NR3A, resulted in smaller NMDA/glycine-evoked currents than NR1/NR2A alone^{5,6,15} (in our case, when examined about 2 d after oocyte injection). This result suggests a dominant interfering effect of NR3 subunits when co-injected with NR1 and NR2 subunits.

We estimate that the EC_{50} (the effective concentration for half-maximum response) for glycine activation of NR1/NR3B receptors is about 5 μ M. Desensitization occurred above 10 μ M glycine, preventing calculation of a more accurate EC_{50} (Fig. 2a). Notably, for NR1/NR3A receptors the EC_{50} was even lower ($\sim$ 1 μ M glycine), with detectable responses at 0.1 μ M, and rapid desensitization at 3 μ M and above (Fig. 2f). We found that D-serine, D-alanine, D-cycloserine and 1-aminocyclopropanecarboxylic acid (ACPC), which are all co-agonists of conventional NMDARs through the glycine-binding site on NR1 (ref. 16), inhibited glycine-induced currents of NR1/NR3B receptors (Fig. 2c). Among these, only D-serine activated NR1/NR3B receptors in the absence of glycine (Fig. 2c), but with lower efficacy than glycine. At NR1/NR3A receptors, D-serine was an even more potent antagonist (Fig. 2h) and did not independently manifest any agonist activity.

The effect of traditional NMDAR antagonists and channel blockers further demonstrated the altered pharmacological properties conferred by the NR3 family of subunits. A competitive antagonist at the glutamate site of NR2 subunits¹⁷, D-(–)-2-amino-5-phosphonopentanoate (AP5), had little effect on glycine-induced currents of NR1/NR3B (Fig. 2d) or NR1/NR3A (Fig. 2h) receptors. In contrast, 5,7-dichlorokynurenic acid (5,7-DCKA), a noncompetitive antagonist at the glycine site of NR1 (ref. 18), virtually abolished glycine-evoked currents of NR1/NR3B receptors (Fig. 2d). Strychnine, an antagonist of inhibitory, chloride-permeable glycine receptors¹⁹, displayed minimal effects (Fig. 2d). NR1/NR3B (Fig. 2e) and NR1/NR3A (Fig. 2h) receptors exhibited dramatically reduced sensitivity to the open-channel blockers Mg^{2+} , MK-801 and memantine when compared with NR1/NR2A receptors^{20–24}.

Next, we examined the voltage dependence and permeation of channels operated by NR1/NR3B receptors. *I–V* curves remained nearly linear between –80 and +40 mV in the presence of 0.5 mM Mg^{2+} (Fig. 3a, b), reflecting relative Mg^{2+} insensitivity at physiologically relevant voltages and further distinguishing NR1/NR3B receptors from conventional NMDARs^{20,21}. However, between –80 and –120 mV, the channel became outwardly rectifying, indicating a shift in the voltage-dependent Mg^{2+} block towards more negative potentials compared with NR1/NR2 receptors (data not shown). Replacement of cations in the external bath with the large cation N-methyl-D-glucamine (NMDG) completely abolished the glycine-induced inward current (Fig. 3a, b), suggesting that NR1/NR3B channels are impermeable to large cations and anions (Cl^-) but permeable to small cations such as Na^+ . Similarly, Mg^{2+} did not

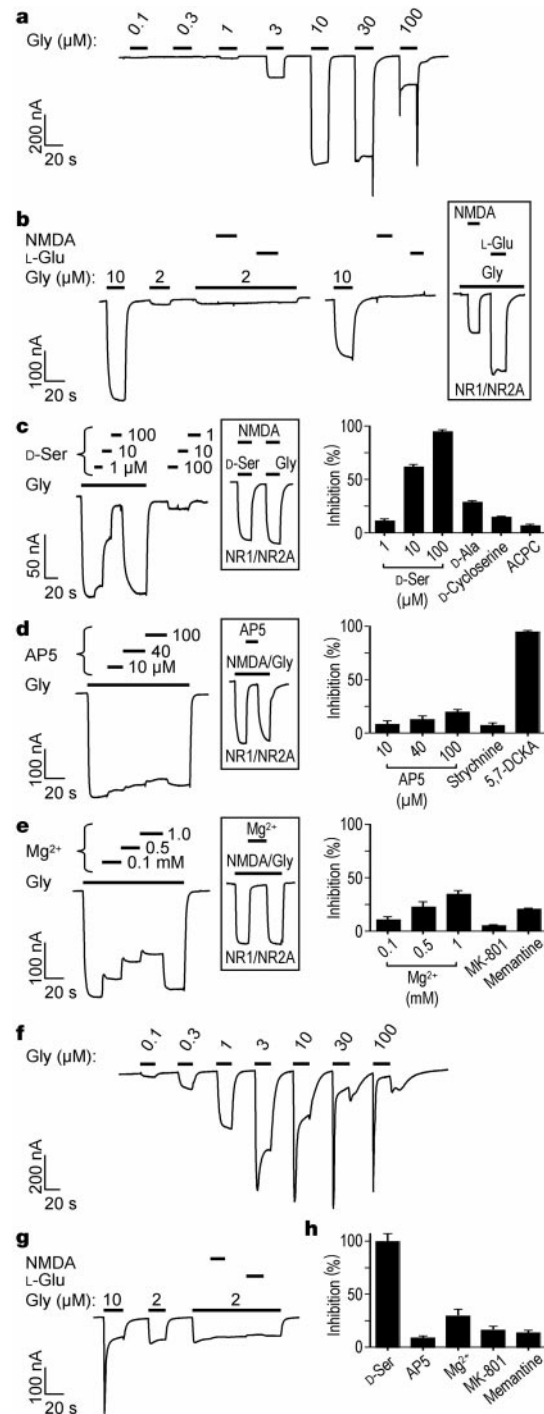


Figure 2 Pharmacological characterization of NR1/NR3B or -3A receptors in *Xenopus* oocytes. Glycine-evoked currents were recorded from oocytes injected with NR1/NR3B (a–e) or NR1/NR3A (f–h). Insets, NMDA/glycine-evoked NR1/NR2A currents for comparison. Glycine and NMDA (or L-glutamate) concentrations were 10 and 100 μ M, respectively, unless otherwise indicated. **a**, Dose-response of glycine-evoked NR1/NR3B currents. **b**, NMDA and L-glutamate did not potentiate glycine-evoked currents, and did not evoke a glycine-independent response. **c**, Left, D-serine evoked small currents alone but inhibited the glycine response; right, inhibition (mean $\pm$ s.e.m.) of glycine-evoked currents by D-serine, D-alanine (30 μ M), D-cycloserine (30 μ M) or ACPC (1 μ M). **d**, Inhibition by AP5, strychnine (10 μ M), and 5,7-DCKA (100 μ M). **e**, Inhibition by Mg^{2+} , MK-801 (10 μ M), and memantine (12 μ M). **f**, Dose response of glycine-evoked NR1/NR3A currents. **g**, NMDA and L-glutamate did not potentiate glycine-evoked currents. **h**, Inhibition of glycine (2 μ M)-evoked currents by D-serine (10 μ M), AP5 (100 μ M), Mg^{2+} (1 mM), MK-801 (10 μ M) and memantine (12 μ M). Data are representative of recordings from 3–9 oocytes in each case.

permeate NR1/NR3B channels at potentials positive to -80 mV (Fig. 3b). Increasing the Ba^{2+} concentration from 1 to 10 mM caused virtually no change in reversal potential for NR1/NR3B receptors, whereas NR1/NR2A receptors underwent a rightward shift of about 20 mV (Fig. 3c, d). Furthermore, with Ca^{2+} in the external bath, activation of NR1/NR3B receptors evoked far less Ca^{2+} -triggered Cl^- current than NR1/NR2 receptors²⁵. These results suggest that NR1/NR3B channels are less permeable to divalent cations such as Ca^{2+} and Ba^{2+} than are conventional NMDARs. Similar results were obtained for channels containing NR3A instead of NR3B.

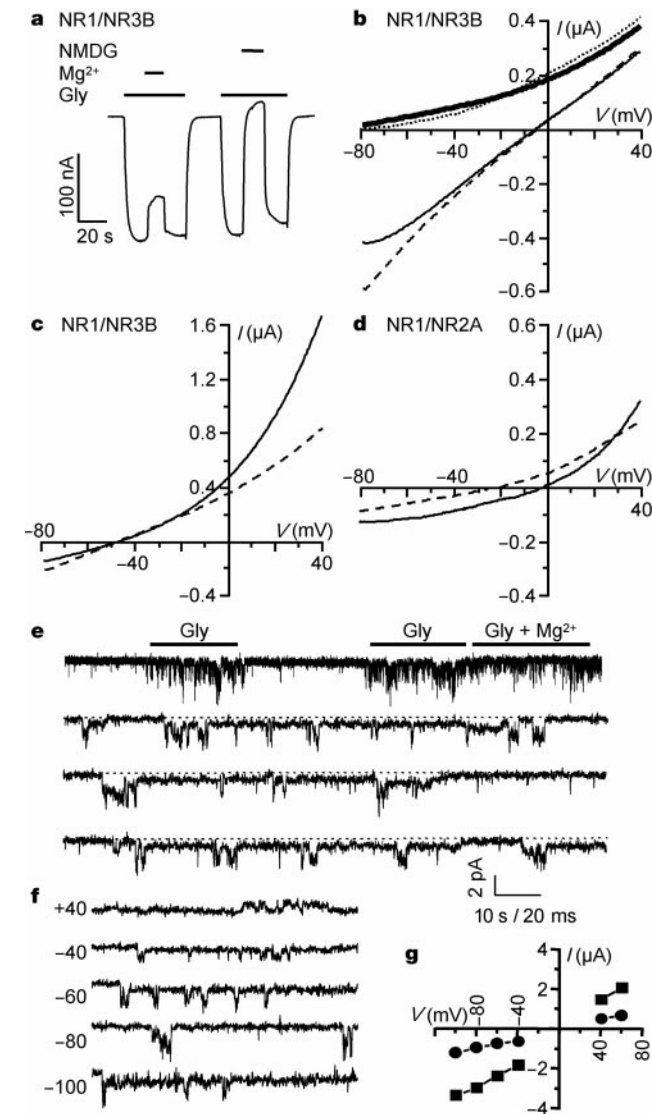


Figure 3 Ion selectivity (a–d) and single-channel analysis (e–g) of NR1/NR3B receptors in *Xenopus* oocytes. **a**, Inhibition of NR1/NR3B currents evoked by glycine ($10 \mu\text{M}$) after addition of Mg^{2+} (0.5 mM) or replacement of cations in the bath solution with 90 mM NMDG. **b**, Glycine-evoked NR1/NR3B currents during voltage ramps (20 mV s^{-1}) in normal bath solution containing 0.5 mM Mg^{2+} (thin solid line) or no added Mg^{2+} (dashed line), and in isosmotic solution containing 90 mM NMDG (with 1 mM Ba^{2+} ; thick solid line) or 10 mM Mg^{2+} (with 2 mM Na^+ , balance NMDG; dotted line). **c**, **d**, NR1/NR3B currents evoked by glycine (**c**) and NR1/NR2A currents evoked by NMDA ($100 \mu\text{M}$) and glycine (**d**) during voltage ramps in 1 mM Ba^{2+} (with 20 mM Na^+ , balance NMDG; dashed line) versus 10 mM Ba^{2+} (with 2 mM Na^+ , balance NMDG; solid line). **e**, Top trace, glycine-activated single-channel currents ($\pm 0.5 \text{ mM}$ Mg^{2+}) in outside-out patches at -60 mV ; lower traces, epochs at higher time resolution. Dotted line indicates zero current level. **f**, Glycine-evoked single-channel currents at various membrane potentials. **g**, Current–voltage relationships of main conductance (squares) and subconductance state (circles) from total amplitude histogram plots ($n = 5$, s.e.m. within size of symbols).

Single-channel recording of NR1/NR3B channels from outside-out patches of oocyte membrane confirmed the unique properties of these receptors, including glycine activation and reduced sensitivity to Mg^{2+} . From total amplitude histograms fit with gaussian components⁷, channels operated by NR1/NR3B receptors manifested a main unitary conductance of $37 \pm 0.8 \text{ pS}$ and a subconductance of $12 \pm 0.5 \text{ pS}$ ($n = 5$; Fig. 3e). Furthermore, with most anions in the patch electrode replaced by impermeant gluconate, the reversal potentials for both conductance states remained near zero, confirming that the channels were cation selective (Fig. 3f, g). Receptors containing NR3A manifested a similar unitary conductance.

Previous studies of *Xenopus* oocytes co-injected with NR1 and NR2A have reported small, glycine-induced inward currents⁴. However, glutamate or NMDA potentiated these glycine-evoked currents and Mg^{2+} or AP5 inhibited them, thus distinguishing them from those reported here. Formation of functional channels in oocytes injected with NR1 alone¹⁴ has been attributed to co-assembly of NR1 and XenU1, a unitary glutamate receptor subunit endogenously expressed by *Xenopus* oocytes²⁶. Although glycine-evoked currents observed in NR1/NR3-injected oocytes could also result from co-assembly with XenU1, this explanation is unlikely for the following reasons. First, XenU1 contains a glutamate-binding site²⁷, and receptors expressed by NR1-injected oocytes (NR1/XenU1 receptors) require glutamate or NMDA in addition to glycine for full activation¹⁴. Receptors composed of an NR3 subunit in addition to NR1 and XenU1 subunits would presumably retain this requirement for a glutamate or NMDA co-agonist. Second, MK-801, Mg^{2+} and AP5 inhibit NR1/XenU1 receptors expressed by NR1-injected oocytes¹⁴, whereas receptors expressed by NR1/NR3-

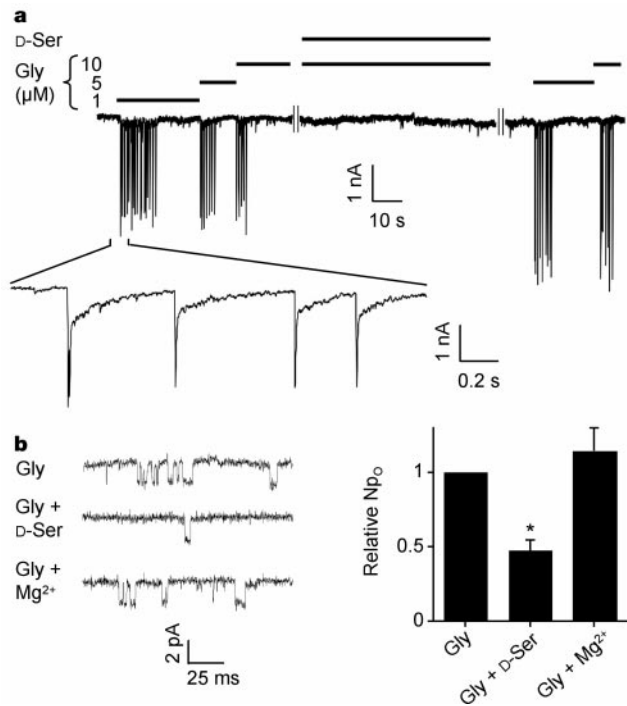


Figure 4 Physiological responses to glycine in rat cerebrocortical neurons. **a**, Whole-cell recording from cultured neurons in the presence of strychnine ($10 \mu\text{M}$) revealed glycine-evoked bursts of action currents. The glycine-evoked response was inhibited by D -serine ($10 \mu\text{M}$). **b**, Single-channel currents from outside-out patches of cerebrocortical neurons in response to $2.5 \mu\text{M}$ glycine at -60 mV manifested a main conductance state of $38 \pm 1.0 \text{ pS}$ ($n = 3$ of 12 patches recorded for 5–30 min). Channel activity was decreased by D -serine ($100 \mu\text{M}$) (asterisk, $P < 0.01$, Student's t -test), but 1 mM Mg^{2+} had little effect. Np_0 represents the product of the number (N) of channels in the patch and the single-channel open probability (p_0).

injected oocytes exhibited drastically reduced sensitivity to these antagonists. Regardless of whether *Xenopus* oocytes provide an additional subunit to the NMDAR complex, NR3 subunits dramatically alter the physiological and pharmacological properties of receptors formed when co-injected with NR1, and imbue them with the properties of an excitatory glycine receptor.

To investigate the physiological relevance of the NR3 family, we monitored whole-cell currents from cultured cerebrocortical neurons dissociated on P0 (recorded at days 8–13 *in vitro*) or E16 (recorded at day 19 *in vitro*), at which times NR3A is maximally expressed and NR3B may also be present⁷. In 22 of 26 recordings performed in the presence of strychnine, glycine (1–10 μ M) triggered neuronal bursting, accompanied by little inward current monitored at the cell body; D-serine inhibited this effect in 5 of 22 glycine-responsive cells (Fig. 4a). Thus, the pharmacological profile matched that of NR1/NR3A or -3B receptors in about one-quarter of the recordings. If conventional NMDARs had been involved (for example, if submaximal glycine were present in conjunction with endogenous glutamate), then D-serine would have enhanced the responses rather than inhibiting them, as observed in other recordings. We ascribed this glycine-triggered bursting phenomenon to a presynaptic or network effect, as it was inhibited by tetrodotoxin. Whatever the mechanism, these glycine-mediated responses seem to affect the threshold for bursts of action potentials within a network, and a substantial number of these responses displayed the pharmacology of NR1/NR3A or -3B receptors. Additionally, application of glycine to outside-out patches formed from the cell body of cerebrocortical neurons evoked single channels of appropriate size (38 pS) and with appropriate pharmacology for the receptors we describe: in the presence of strychnine, D-serine inhibited the glycine-activated channels and Mg²⁺ had little effect (Fig. 4b). The easily desensitized nature of these extrasynaptic currents, similar to those of recombinant NR1/NR3A channels, may have obscured them from previous workers.

In vivo, co-assembly of NR3A or NR3B with NR1 and possibly NR2 subunits may impart to heteromeric NMDARs some or all of the unique properties described here, and may account for the observed diversity in NMDAR function^{28–30}. NR3 may also be physiologically important during synapse formation and long-term potentiation when NMDARs are thought to be functionally silent in the absence of AMPA receptors. In this case, insensitivity of NR3-containing channels to Mg²⁺ may obviate the need for AMPAR-mediated depolarization. Rather, NR1/NR3-mediated depolarization may activate conventional NMDARs at the synapse. For these reasons, standard pharmacological methods that have been used to study NMDAR activity in culture and *in vivo* must be re-evaluated in light of NR3A or NR3B expression. □

Methods

Cloning of cDNA

Using degenerate polymerase chain reaction (PCR)⁵, we amplified cDNA fragments with homology to glutamate receptor family members from rat brain mRNA. Two larger cDNA fragments (5-2 clone 1 and 5-2 clone 2) were obtained by screening a rat brain cDNA library using probes constructed from the PCR products. Four additional partial cDNA clones were also obtained by screening a Rapid-Screen rat brain cDNA library panel (OriGene Technologies) using two pairs of more specific, nested primers (f4, TGCTGCTATGGCTACTGTCATC; r4, ATGACAGCAGCCAGGTGGCCGT; f5, CACACATGGCTGTGACCAGC; and r5, AGAATGGCATAGCACAGGTG). The 2847-bp *XhoI*–*XbaI* fragment of one of these clones (B4) was ligated to the 345-bp *EcoRI*–*XhoI* fragment of 5-2 clone 2 in pCMV6-XL4 to generate full-length NR3B cDNA.

In situ hybridization

We prepared 20- μ m sections from brain and spinal cord of adult Sprague–Dawley rats as described⁶. Antisense and sense (control) cRNA probes, corresponding to the ~400-bp *AvrII*–*BamHI* fragment of NR3B, were labelled by *in vitro* transcription using [³³P]UTP (NEN) and T3 or T7 RNA polymerase (Roche) or the digoxigenin (DIG) RNA labelling kit (Roche). Hybridization and washing were performed as described⁶, except that 200 μ l of ³³P-labelled (10⁸ counts per min (c.p.m.) per ml at 10³ c.p.m. ng⁻¹) or DIG-labelled (100 μ g ml⁻¹) cRNA were used for hybridization. DIG labelling was detected using alkaline phosphatase-labelled sheep anti-DIG Fab fragments and NBT/BCIP (Roche).

Immunocytochemistry

Sections dried overnight were washed in PBS and treated with 40% (v/v) methanol, 3% H₂O₂ and 0.3% Triton X-100 buffer in PBS for 20 min. Sections were then blocked with 3% horse serum plus 0.3% Triton X-100 and incubated with the anti-NR3A monoclonal antibody K35/40 diluted 1:5, followed by biotinylated secondary antibody (Vector, 1:200), avidin-biotin-peroxidase solution (Vector, 1:100), and SIGMA FAST 3,3'-diaminobenzidine with the Metal Enhancer Tablet Set (Sigma).

Electrophysiological recording

For expression in *Xenopus* oocytes, NR3B was subcloned into pGEM-HE modified to contain a T3 site upstream from the 5' untranslated region (UTR), and *AscI*, *PacI* and *PmeI* sites downstream from the 3' UTR. NR1, NR2A, NR3A and NR3B cRNAs were produced using mMessage mMachine kits (Ambion). Oocytes were co-injected with 2 ng NR1, 4 ng NR2A and/or 4–12 ng NR3B (or NR3A) cRNAs. Recordings were performed 2–7 d after injection. Unless otherwise indicated, two-electrode voltage-clamp recordings were performed at –80 mV in bath solution containing (in mM) 90 NaCl, 1 KCl, 1.5 BaCl₂ and 10 HEPES buffer (pH 7.5)⁶. Single-channel recording was performed on outside-out patches from oocytes⁷. The intracellular (patch-pipette) solution contained (in mM) 90 sodium gluconate, 10 NaCl, 2 ATP, 0.25 GTP, 10 BAPTA and 10 HEPES (pH 7.3). The extracellular solution contained (in mM) 100 NaCl, 2.5 KCl, 2 BaCl₂ and 5 HEPES (pH 7.35). For recordings from cerebrocortical neurons, cultures were prepared and whole-cell and single-channel recordings were performed as described^{7,23}, except that 5–10 μ M strychnine was generally added to the bathing medium.

Received 29 October 2001; accepted 14 January 2002.

Published online 30 January 2002, DOI 10.1038/nature715.

- Lipton, S. A. & Rosenberg, P. A. Excitatory amino acids as a final common pathway for neurologic disorders. *N. Engl. J. Med.* **330**, 613–622 (1994).
- Cull-Candy, S., Brickley, S. & Farrant, M. NMDA receptor subunits: diversity, development and disease. *Curr. Opin. Neurobiol.* **11**, 327–335 (2001).
- Monyer, H. *et al.* Heteromeric NMDA receptors: molecular and functional distinction of subtypes. *Science* **256**, 1217–1221 (1992).
- Meguro, H. *et al.* Functional characterization of a heteromeric NMDA receptor channel expressed from cloned cDNAs. *Nature* **357**, 70–74 (1992).
- Ciabarra, A. M. *et al.* Cloning and characterization of χ -1: a developmentally regulated member of a novel class of the ionotropic glutamate receptor family. *J. Neurosci.* **15**, 6498–6508 (1995).
- Sucher, N. J. *et al.* Developmental and regional expression pattern of a novel NMDA receptor-like subunit (NMDAR-L) in the rodent brain. *J. Neurosci.* **15**, 6509–6520 (1995).
- Das, S. *et al.* Increased NMDA current and spine density in mice lacking the NMDA receptor subunit NR3A. *Nature* **393**, 377–381 (1998).
- Hirai, H., Kirsch, J., Laube, B., Betz, H. & Kuhse, J. The glycine binding site of the N-methyl-D-aspartate receptor subunit NR1: identification of novel determinants of co-agonist potentiation in the extracellular M3–M4 loop region. *Proc. Natl Acad. Sci. USA* **93**, 6031–6036 (1996).
- Kuryatov, A., Laube, B., Betz, H. & Kuhse, J. Mutational analysis of the glycine-binding site of the NMDA receptor: structural similarity with bacterial amino acid-binding proteins. *Neuron* **12**, 1291–1300 (1994).
- Wafford, K. A. *et al.* Identification of amino acids in the N-methyl-D-aspartate receptor NR1 subunit that contribute to the glycine binding site. *Mol. Pharmacol.* **47**, 374–380 (1995).
- Anson, L. C., Chen, P. E., Wyllie, D. J., Colquhoun, D. & Schoepfer, R. Identification of amino acid residues of the NR2A subunit that control glutamate potency in recombinant NR1/NR2A NMDA receptors. *J. Neurosci.* **18**, 581–589 (1998).
- Laube, B., Hirai, H., Sturgess, M., Betz, H. & Kuhse, J. Molecular determinants of agonist discrimination by NMDA receptor subunits: analysis of the glutamate binding site on the NR2B subunit. *Neuron* **18**, 493–503 (1997).
- Pérez-Otaño, I. *et al.* Assembly with the NR1 subunit is required for surface expression of NR3A-containing NMDA receptors. *J. Neurosci.* **21**, 1228–1237 (2001).
- Moriyoshi, K. *et al.* Molecular cloning and characterization of the rat NMDA receptor. *Nature* **354**, 31–37 (1991).
- Nishi, M., Hinds, H., Lu, H.-P., Kawata, M. & Hayashi, Y. Motoneuron-specific expression of NR3B, a novel NMDA-type glutamate receptor subunit that works in a dominant-negative manner. *J. Neurosci.* **21**, RC185 (2001).
- McBain, C. J., Kleckner, N. W., Wyrick, S. & Dingledine, R. Structural requirements for activation of the glycine coagonist site of N-methyl-D-aspartate receptors expressed in *Xenopus* oocytes. *Mol. Pharmacol.* **36**, 556–565 (1989).
- Davies, J., Francis, A. A., Jones, A. W. & Watkins, J. C. 2-Amino-5-phosphonovalerate (2APV), a potent and selective antagonist of amino acid-induced and synaptic excitation. *Neurosci. Lett.* **21**, 77–81 (1981).
- Baron, B. M. *et al.* Activity of 5,7-dichlorokynurenic acid, a potent antagonist at the N-methyl-D-aspartate receptor-associated glycine binding site. *Mol. Pharmacol.* **38**, 554–561 (1990).
- Curtis, D. R., Hosli, L. & Johnston, G. A. Inhibition of spinal neurons by glycine. *Nature* **215**, 1502–1503 (1967).
- Mayer, M. L., Westbrook, G. L. & Guthrie, P. B. Voltage-dependent block by Mg²⁺ of NMDA responses in spinal cord neurones. *Nature* **309**, 261–263 (1984).
- Nowak, L., Bregestovski, P., Ascher, P., Herbert, A. & Prochiantz, A. Magnesium gates glutamate-activated channels in mouse central neurons. *Nature* **307**, 462–465 (1984).
- Wong, E. H. *et al.* The anticonvulsant MK-801 is a potent N-methyl-D-aspartate antagonist. *Proc. Natl Acad. Sci. USA* **83**, 7104–7108 (1986).
- Chen, H.-S. V. *et al.* Open-channel block of N-methyl-D-aspartate (NMDA) responses by memantine: therapeutic advantage against NMDA receptor-mediated neurotoxicity. *J. Neurosci.* **12**, 4427–4436 (1992).
- Bormann, J. Memantine is a potent blocker of N-methyl-D-aspartate (NMDA) receptor channels. *Eur. J. Pharmacol.* **166**, 591–592 (1989).
- Leonard, J. P. & Kelso, S. R. Apparent desensitization of NMDA responses in *Xenopus* oocytes involves calcium-dependent chloride current. *Neuron* **4**, 53–60 (1990).

26. Soloviev, M. M. & Barnard, E. A. *Xenopus* oocytes express a unitary glutamate receptor endogenously. *J. Mol. Biol.* **273**, 14–18 (1997).
27. Ishimaru, H. et al. A unitary non-NMDA receptor short subunit from *Xenopus*: DNA cloning and expression. *Receptors Channels* **4**, 31–49 (1996).
28. von Euler, M., Li-Li, M., Whittemore, S., Seiger, A. & Sundström, E. No protective effect of the NMDA antagonist memantine in experimental spinal cord injuries. *J. Neurotrauma* **14**, 53–61 (1997).
29. Paleček, J., Abdrachmanova, G., Vlachová, V. & Vyklíček, L. Jr Properties of NMDA receptors in rat spinal cord motoneurons. *Eur. J. Neurosci.* **11**, 827–836 (1999).
30. Souverbie, E., Mo, L.-L., Liu, Y., von Euler, G. & Sundström, E. Pharmacological characterization of [³H]MK-801 binding in the rat spinal cord. *Eur. J. Pharmacol.* **307**, 347–353 (1996).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com>).

Acknowledgements

We thank H.-S. V. Chen for discussions and T. Ciabarra and M. Forcina for preliminary work. Monoclonal antibody K35/40 was made in collaboration with J. S. Trimmer and N. J. Sucher. NR1 and NR2A cDNAs were gifts of S. F. Heinemann and S. Nakanishi, respectively. This work was supported in part by grants from the National Institutes of Health. S.A.L. and D.Z. are co-senior authors.

Competing interests statement

The authors declare competing financial interests: details accompany the paper on Nature's website (<http://www.nature.com>).

Correspondence and requests for materials should be addressed to S.A.L. (e-mail: slipton@burnham.org) or D.Z. (e-mail: dzhang@burnham.org). The rat NR3B sequence has been deposited in GenBank under accession number AF440691.

Establishment of developmental precision and proportions in the early *Drosophila* embryo

Bahram Houchmandzadeh^{†‡}, Eric Wieschaus^{*} & Stanislas Leibler^{*‡§}

^{*} Howard Hughes Medical Institute, Department of Molecular Biology, Princeton University, Princeton, New Jersey 08544, USA

[†] CNRS, Laboratoire de Spectrométrie Physique, BP87, 38402, St-Martin D'Heres Cedex, France

[‡] Department of Physics, Princeton University, Princeton, New Jersey 08544, USA

During embryonic development, orderly patterns of gene expression eventually assign each cell in the embryo its particular fate. For the anteroposterior axis of the *Drosophila* embryo, the first step in this process depends on a spatial gradient of the maternal morphogen Bicoid (Bcd). Positional information of this gradient is transmitted to downstream gap genes, each occupying a well defined spatial domain^{1–4}. We determined the precision of the initial process by comparing expression domains in different embryos. Here we show that the Bcd gradient displays a high embryo-to-embryo variability, but that this noise in the positional information is strongly decreased ('filtered') at the level of *hunchback* (*hb*) gene expression. In contrast to the Bcd gradient, the *hb* expression pattern already includes the information about the scale of the embryo. We show that genes known to interact directly with Hb are not responsible for its spatial precision, but that the maternal gene *staufen* may be crucial.

Development is a precise process. All biochemical phenomena are prone to variation; at each level, correcting mechanisms should exist to avoid the transmission of errors and their amplification to downstream genes. The establishment of a morphogen gradient is

a good example of an error-prone process. The morphogen gradient model¹ is the most widely accepted model describing how cells in early embryos acquire their positional information: because the concentration of a morphogen decreases with the distance from one pole, measuring this concentration informs a cell of its position inside the embryo. In *Drosophila* embryos, the maternal gene *bcd* possesses many properties of a prototypical morphogen^{2–4}. In such models, downstream genes, such as *hb*, are activated by a 'threshold' mechanism: only cells that measure concentrations of Bcd above a certain level switch on the *hb* gene.

It is hard to imagine how precision could be achieved in this simple threshold model. To produce exactly the same concentration profile of Bcd in each embryo, the mother would have to control the exact amount of messenger RNA deposited in each embryo, its localization, and the amount of protease responsible for the morphogen degradation. Any error in these parameters modifies the morphogen gradient and induces error in the positional information delivered to downstream genes⁵. Even if these parameters were precisely controlled, the gradient profile would still be sensitive to environmental conditions such as temperature. Another limitation of a simple gradient is conservation of proportions: an exponential gradient has its own length scale λ (depending on the diffusion coefficient D and the protein degradation rate ω , $\lambda^2 = D/\omega$), which is independent of the length of the embryo.

We address here both the issue of the error in the positional information of the Bcd gradient and the establishment of the spatial proportions in the embryo. To quantify the precision of gene expression, we measured protein profiles at cycle 14 by immunofluorescence staining (Fig. 1, see Methods). The Bcd protein profiles in about 100 wild-type embryos during early cycle 14 (before significant membrane invagination) are shown in Fig. 2a. The profile displays a high embryo-to-embryo variability. To quantify this variability, we measured the position (x_{Bcd}) along the embryo at which each curve crosses a certain threshold (t), chosen here as 0.23 of the maximal intensity (see Methods). These positions are spread over 30% of the embryonic length (EL) (Fig. 2b), and have a standard deviation (σ_{Bcd}) of 0.07 EL. This means that the positional error of the Bicoid gradient is greater than five nuclei in 50% of embryos. Another way to quantify the variability, which is not sensitive to the normalization of the Bcd protein profile, is to measure the slope of the exponential decay of each curve, λ (Fig. 2c). The standard deviation of λ is 0.045 EL, which corresponds to the

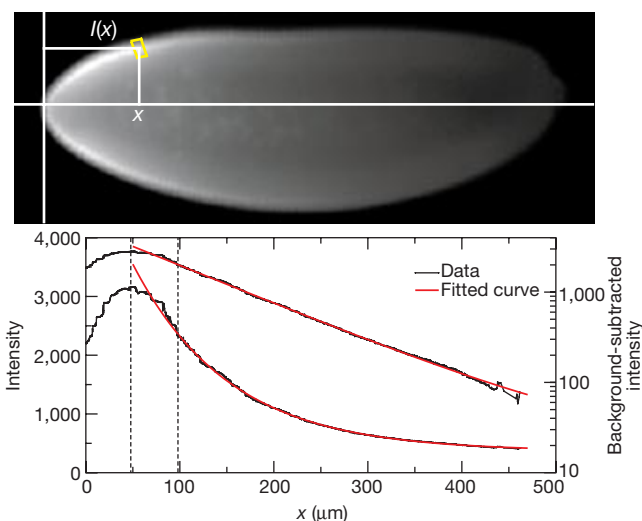


Figure 1 Typical image of Bcd staining, and the profile of the extracted curve. For Bcd, profiles are fitted well by an exponential decay. Right axis, intensity in a linear scale; left axis, background-subtracted intensity in log scale. The size of the area used to compute the intensity is exaggerated (see Methods).

[§] Present address: Laboratory of Living Matter, Rockefeller University, 1230 York Avenue, Box 34, New York, New York 10021, USA

same variability we measured for x_{Bcd} ($\sigma_{\text{Bcd}} = \sigma_{\lambda} \ln(1/t)$).

In contrast to Bcd, the Hb protein profile displays an extreme reproducibility from embryo to embryo. The Hb profile in about 100 embryos from early to late cycle 14 is shown in Fig. 2d. We quantified the *hb* distribution using the point (x_{Hb}) at which each profile crosses the 0.5 threshold (Fig. 2b). The standard deviation of x_{Hb} (σ_{Hb}) is 0.01 EL, meaning that two-thirds of embryos have Hb boundaries defined more precisely than the size of one nucleus. The information about the embryo scale is also revealed at *hb* expression level. As discussed above, the Bicoid exponential profile should not be affected by embryo length. When x_{Bcd} is plotted against EL (Fig. 2e), the correlation coefficient is indeed negligible ($P = 0.40$). A similar lack of correlation is observed between the values of λ and EL. In contrast to Bcd, however, the position of the Hb boundary displays a strong correlation with the embryo length (Fig. 2f). The linear (r) and Spearman's rank (r_s) correlation coefficients between x_{Hb} and EL are 0.84 and 0.82, respectively ($P < 10^{-20}$; all P -values are computed for r_s). *hb* mRNA displays the same precision and conservation of proportions as Hb protein. The spatial position of the *hb* mRNA boundary in early cycle 14 embryos has a standard deviation of 0.01 EL, and displays a high correlation with the egg length ($r_s = 0.88$, $P < 10^{-8}$).

The precision of the Hb boundary compared with the variability of the Bcd gradient could seem at odds with experiments where Bcd dosage has been modified. However, when the Hb boundary position as a function of Bcd dosage is compared with the expected value from a simple threshold model (Table 1), the measured shift is significantly smaller than that expected, even when the reduced efficiency we measure for the *bcd* transgenes is taken into consid-

Table 1 Hb boundaries in different *bcd* backgrounds

Background	Measured	Expected if efficiency is 50%	Expected if efficiency is 100%
x_{Hb} (<i>bcd</i> 1 ×)	0.41	NA	0.30
x_{Hb} (<i>bcd</i> wild type)	0.49	NA	NA
x_{Hb} (<i>bcd</i> 4 ×)	0.56	0.60	0.68
x_{Hb} (<i>bcd</i> 6 ×)	0.59	0.68	0.79

The measured and expected values for the position of the Hb boundary are shown for different *bcd* dosages. The expected shift of Hb boundary is $\lambda \ln(n/n_0)$, where n is the number of copies of *bcd* compared with the wild type ($n_0 = 2$). The σ_{Hb} is 1% for the wild-type embryos and *bcd* 1 ×, 1.5% for *bcd* 4 ×, and 2% for *bcd* 6 ×. The statistical significance of the differences between the expected and measured x_{Hb} , $P < 10^{-16}$. For 4 × and 6 ×, flies carrying two copies of a *bcd* transgene on chromosome X were used. Heterozygous mothers have four copies of *bcd* (*bcd* 4 ×) and homozygous ones possess six copies (*bcd* 6 ×). For 1 ×, mothers heterozygous for *bcd*²¹ were used. *bcd* wild-type flies are 2 ×. NA, not applicable.

eration. The shift cannot be explained quantitatively if *hb* were activated only by Bicoid.

In mid-embryo, the Hb concentration decreases from the highest to the lowest value across about 0.1 EL. The corresponding change in Bcd concentration in this region is only 30% (corresponding to $1 - \exp(-0.1/0.27)$). If Bcd were the only source of cooperative activation of *hb*, this small change in the Bcd concentration would necessitate a Hill coefficient of more than 10. Such strong cooperative activation would be very sensitive to temperature variations⁶. To measure the temperature sensitivity of the Hb profile, we collected embryos for 1 h at 25 °C, and then allowed them to reach cycle 14 at different temperature (9–29 °C). The developmental time varies strongly as a function of temperature, and ranges from 2 h at 25 °C to 20 h at 9 °C. The induced changes in the Bcd and Hb profiles are shown in Fig. 3c and d. The Bcd profile is sensitive to temperature: this could be expected from a simple diffusing

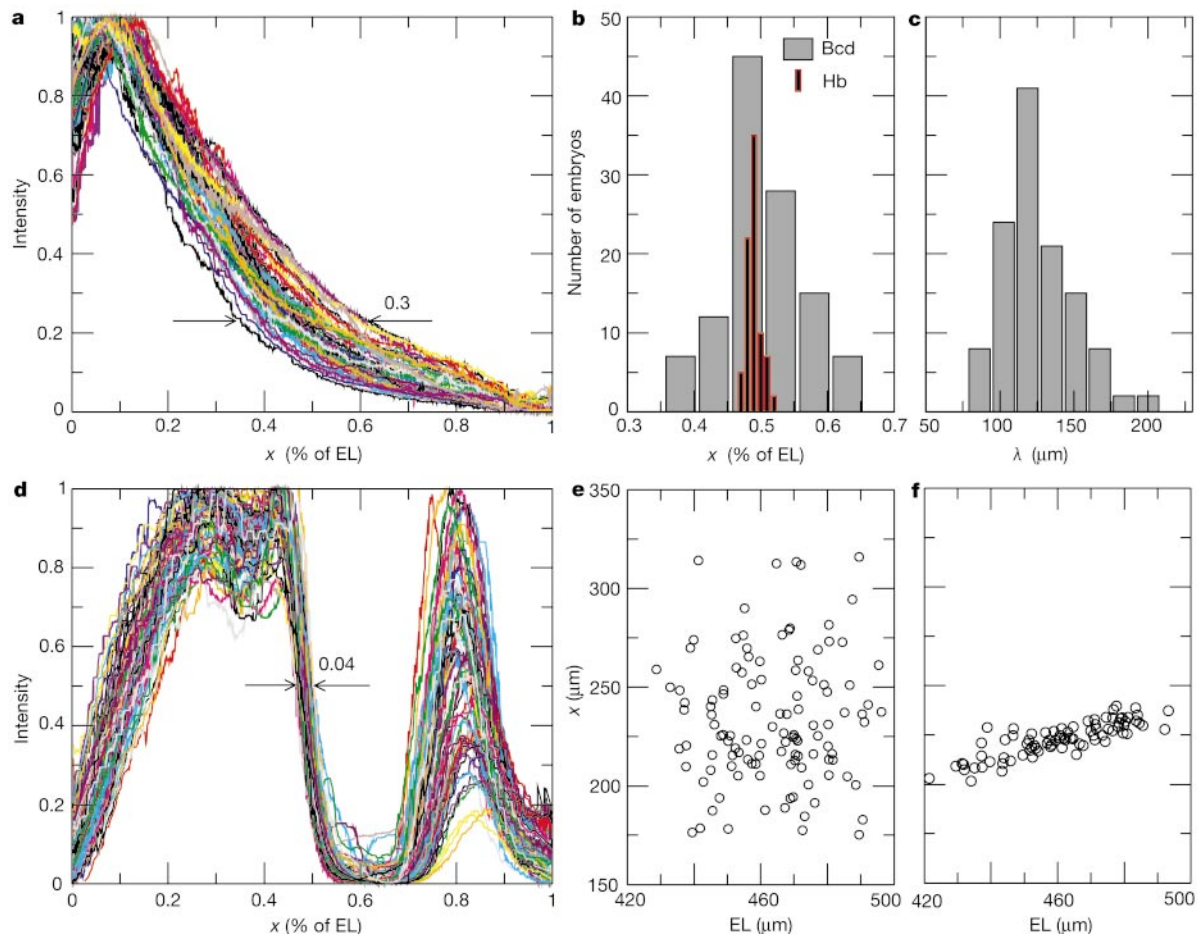


Figure 2 Positional information of Bcd and Hb gradients. **a**, Bcd gradient in about 100 embryos. **b**, Distribution of positions at which each gradient crosses a given threshold: 0.23 for Bcd, 0.5 for Hb. **c**, Distribution of slope of exponential decay for each Bcd profile.

d, Hb gradient in about 100 embryos. **e**, **f**, Position at which each gradient crosses the given threshold versus embryo length (EL) for Bcd (**e**) and Hb (**f**).

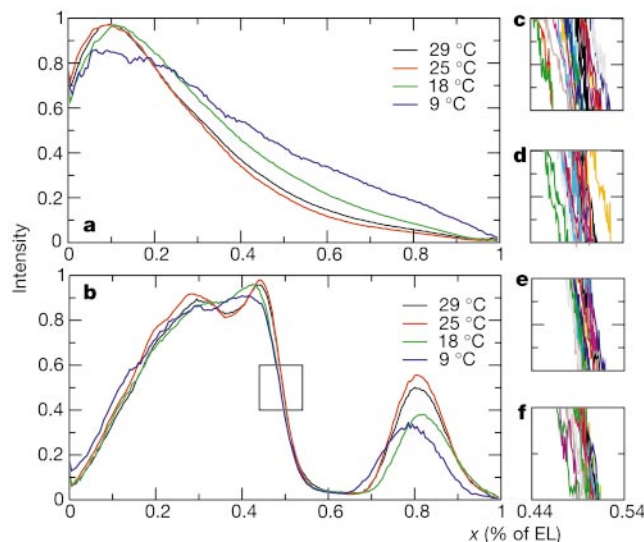


Figure 3 Influence of temperature variation on Bcd and Hb gradients. **a**, Average profiles of Bcd gradients at different temperatures (9–29 °C). **b**, The corresponding average profiles of Hb gradients. **c–f**, Detail of individual Hb profiles at each temperature.

Coordinates of the frame correspond to the rectangle shown in **b**. **c**, 9 °C, $\sigma = 0.016$; **d**, 18 °C, $\sigma = 0.013$; **e**, 25 °C, $\sigma = 0.01$; **f**, 29 °C, $\sigma = 0.011$. The average profiles are computed from 18–32 curves.

morphogen model, in which the protease rate, and thus λ , should strongly depend on temperature. However, no change in the position and variability of Hb was observed. Thus, variations due to temperature changes are also compensated at the level of *hb* expression.

To determine when Hb precision arises, we measured gene expression patterns of *bcd* and *hb* in the early embryo from cycle 12 to 14. No significant change was observed in the Bcd pattern during this period: its variability remained extremely high ($\sigma_{\text{Bcd}} = 0.07$ EL) and its amplitude remained constant throughout early cycle 14. The first trace of zygotic Hb protein was detected at cycle 13, where the amplitude of the signal increased by a factor of 1.5 above that achieved in unfertilized eggs or in embryos at cycle 11 or 12. At cycle 14, the relative amplitude (compared with cycle 12) of the signal is 3.7 (see Methods for fluorescence quantification). The variability of Hb boundary position (σ_{Hb}) during cycle 13 is 0.015 EL, only slightly higher than that observed during cycle 14, and clearly lower than that of Bcd at any stage in development. The scaling of the Hb boundary position relative to the egg length also appears at cycle 13: no significant correlation existed between the Hb boundary and the size of the egg at cycle 12 ($r_s = 0.3$, $P = 0.1$). At cycle 13, r_s rose to 0.7 ($P = 0.00015$), and further increased to 0.8 during cycle 14.

All the genes downstream of *hb* that we have measured (*Kr*, *kni*, *gt*, *eve*) display the same spatial precision as *hb*. We have used various *Drosophila* mutants to identify additional genes involved in the fine regulation of Hb. Our search assumed that the removal of such genes would increase the variability of Hb to reflect the variability of Bcd itself (note that some genes such as *gt* induce a shift in Hb position, but not in its variability). The best candidate for a maternal posterior gradient is *nanos* (*nos*). The removal of *nos* induces a slight shift in Hb boundary, but the variability at this new position is not significantly changed (Table 2). Moreover, the simultaneous removal of maternal *hb* and *nos* cancels this shift. Hence, the induced shift by *nos* can be attributed to auto-activation of *hb*^{7,8}. The removal of *nos* does not abolish the scaling property, and the Hb boundary and EL remain tightly correlated ($r = 0.71$, $r_s = 0.64$, $P < 10^{-5}$). Thus *nos* mutations can modify the average position of the Hb boundary without affecting its precision or scaling.

The natural candidate for zygotic control of Hb is Kruppel (*Kr*). Double staining for these two gene products shows a strong

correlation between the position of their boundaries ($r = 0.90$, $r_s = 0.89$, $P < 10^{-20}$), and regulation of *Kr* by Hb has been previously demonstrated⁹. Repression of *hb* expression by *Kr* has also been reported¹⁰. However, Hb boundary position and its low embryo-to-embryo variability are not affected in *Kr* mutants (Table 2). These results hold even if we consider very late cycle 14 embryos in which other members of gap gene class are active. They seem not to be responsible for the precision of the Hb boundary (Table 2). To test the existence of another, unidentified gene contributing to this process, we removed genes corresponding to about 80% of the *Drosophila* genome by using compound chromosomes, but we observed no effect on the precision of the Hb boundary.

The auto-activation of *hb* by its gene products can be important for the sharpness of the Hb boundary and its average position, but this mechanism alone cannot provide any spatial clue for the precision and scaling. To further investigate this issue, we used embryos homozygous for the *hb*^{6N} mutation. These embryos make detectable protein, even though phenotypically the alleles behave as null^{11,12}. The boundary of gene expression for these embryos is shifted towards the anterior, reflecting a role for Hb in amplification of its own expression, presumably by the P1 enhancer⁸. The precision and scaling of the Hb boundary in this background,

Table 2 Variability of the Hb boundary

Mutation	X_{Hb}	σ	n
Wild type	0.49	0.010	110
<i>oskar</i> ⁶	0.52	0.016	40
<i>nos</i> ^{BN}	0.54	0.016	16
<i>hb</i> ^{mat} <i>nos</i> ^{BN}	0.47	0.014	31
<i>exu</i> ^{PJ42}	0.51	0.020	21
<i>swa</i> ¹	0.52	0.025	17
<i>torso</i> ^{PM51}	0.50	0.014	30
<i>spg</i>	0.51	0.017	9
<i>Kr</i> ¹	0.49	0.011	29
<i>kni</i> (del)	0.48	0.010	11
<i>gt</i> ^{YAB2}	0.46	0.010	16
<i>hb</i> ^{6N}	0.45	0.011	17
X ⁻	0.43	0.014	40
2R ⁻	0.49	0.015	19
2L ⁻	0.47	0.014	16
3L ⁻	0.50	0.020	39

The average position of the Hb boundary in different zygotic and maternal backgrounds is shown. Anterior = 0; posterior = 1. n , number of embryos sampled. Compound chromosomes were used to generate embryos deficient for the X chromosome and individual autosomal arms¹⁸.

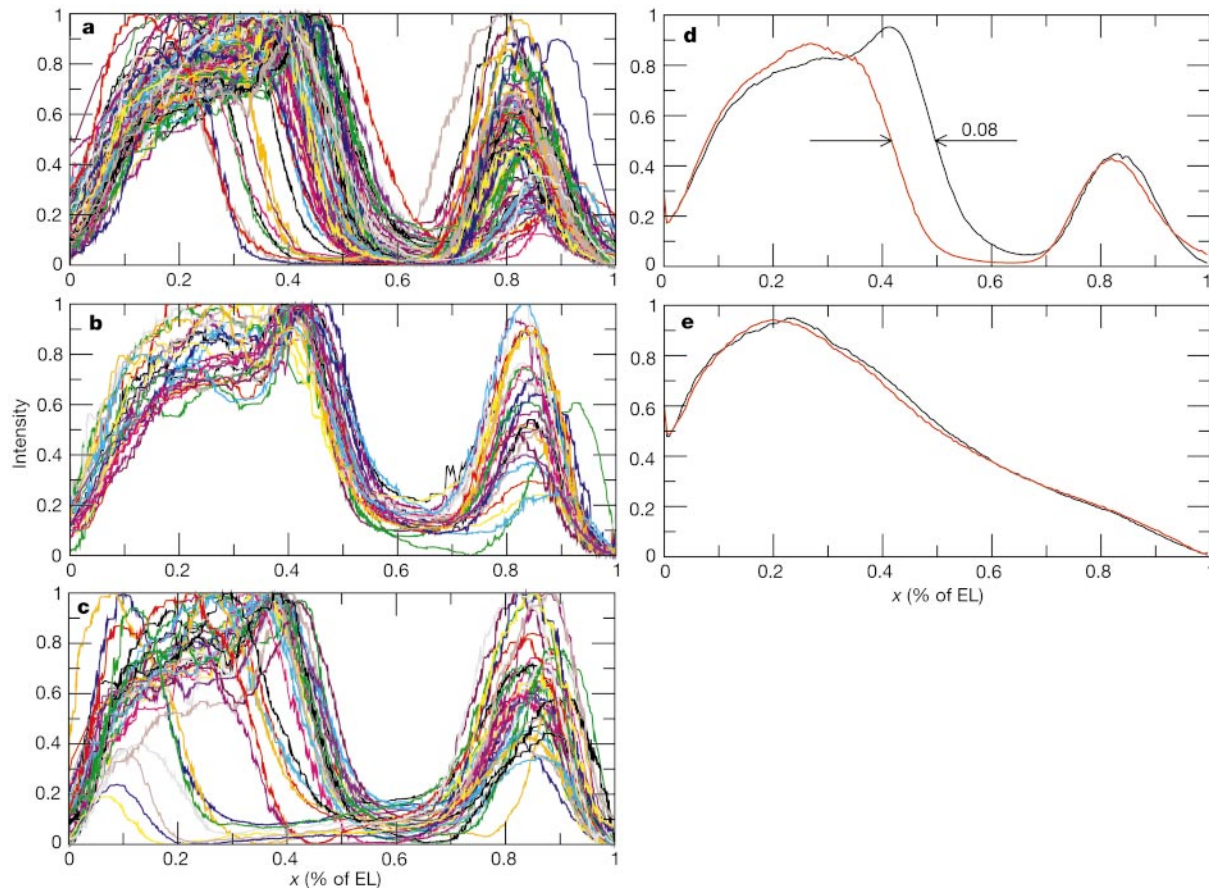


Figure 4 Hb profiles in *stau* background. **a–c**, Hb profiles for about 100 embryos in *stau*^{HL} (**a**), *stau*^{D3} (**b**) and *stau*⁹ (**c**). **d**, Average profile of Hb gradient in *stau*^{HL} background, for populations shifted forwards and backwards of the average. **e**, The

corresponding average Bcd profile for these populations. The average maximum Bcd intensities for these two populations are $1,800 \pm 80$ ($\pm \sigma$) and $1,750 \pm 100$.

however, is equal to that of the wild type.

Among all the mutations we have studied, the only ones that affect Hb boundary precision are certain alleles of the maternal gene *stau*fen (*stau*). In embryos from mothers homozygous for either *stau*^{HL} or *stau*⁹, the Hb boundary position shows a variability of 6%, comparable to the observed Bcd variability (Fig. 4a, c). Surprisingly, this variability is largely reduced (to 2%) in another strong allele of *stau*, D3 (Fig. 4b). Mutations in *stau* disrupt *bcd* and *osk* RNAs and decrease Bcd protein level about twofold. We tested whether the effect of *stau* on Hb was simply an indirect effect of its variable effect on bicoid. From the pool of embryos in *stau*^{HL} background that were double stained for Bcd and Hb, we selected two populations: one that displayed an anterior Hb boundary shift, and one that displayed a posterior shift (Fig. 4d). The corresponding average Bcd profiles for these two populations are very similar, both in the Bcd level and in its spatial distribution (Fig. 4e). Thus, the observed variability in the Hb boundary position may reflect an activity of *stau*fen independent of *bcd*. The disruption of Hb precision in *stau*^{HL} is transmitted to downstream genes, and is not corrected before gastrulation. For instance, double staining for Hb and Kr (data not shown) shows that the variability of the Kr boundary in the *stau*^{HL} background is similar to that of the Hb boundary. Moreover, the positions of these two boundaries remain tightly correlated, as in the wild type.

By quantitatively analysing the protein profiles of maternal morphogens and zygotic gap genes in numerous wild-type and mutant embryos, we have demonstrated two phenomena that take place in the early *Drosophila* development. First, at a very early stage, noise associated with the maternal gradient of Bcd is filtered out, and at the same time the genetic network, which includes the Hb

gap gene, establishes spatial proportions (scaling) in the embryo. It is potentially significant that *stau*fen, the one gene affecting the process, makes a product that localizes to both poles of the egg^{13,14}. More work is needed to establish the mechanisms that control the spatial scaling and precision. It would be then interesting to investigate whether similar phenomena are present in other developmental processes in *Drosophila* and other organisms. □

Methods

Immunostaining of embryos

Embryos were collected at 25 °C (except when temperature variations were studied), heat fixed and labelled with fluorescent probes following previously published protocols¹⁵. All antibodies used were a gift of J. Reintz and D. Kosman¹⁶.

Image analysis

High-resolution digital images ($1,317 \times 1,015$ pixels, 12 bits per pixel) of stained embryos at the same developmental stage and oriented in a lateral projection were taken. Images were focused at mid-embryo to avoid geometric distortion. Intensity profiles were extracted by sliding a rectangle (the size of a nucleus) perpendicular to the embryo along its edges, and computing the average intensity of its pixels, while projecting the coordinates of its centre on the two (anteroposterior and dorsoventral) axes of the embryo. Two curves, corresponding to dorsal and ventral sides of the embryo, were constituted. For consistency, we compared only dorsal profiles. To normalize the intensity profiles, for each curve the minimum and maximum intensity were computed by the average of the 20 lowest- and highest-intensity pixels (corresponding to the size of one nucleus), and the profiles are linearly mapped to a [0,1] interval. The x-axis (when needed) was normalized by the embryo length.

At mid-embryo, the concentration of Hb drops from a high value (normalized to 1) to a low one (normalized to 0). We therefore chose the threshold 0.5 to quantify the spatial position of this boundary. The same threshold is used for the quantification of other gap genes studied.

To characterize Bcd profiles (which are exponential), we chose the threshold for this protein ($t = 0.23$) such that, on average, $x_{Bcd} = x_{Hb}$. The measured variability of

Bcd does not depend on the particular choice of t , and other values (such as 0.5) give the same results.

Profile quantification

Fluorescence antibody staining can be used to determine the relative concentration of a protein in a given background compared to the wild type, if a statistical approach is taken: the typical standard deviation for the maximum fluorescence intensity is about 20%. For intensity calibration, we used 20–30 embryos, so the average maximum intensity displays a standard error of 4% ($0.2/\sqrt{25}$). When a background was compared to wild-type embryos, both mutant and wild-type embryos were stained at the same time and in the same conditions. We used Student's t -test to decide whether two averages were different. The position at which the Bcd gradient crosses a given threshold is subject to an error in the normalization process, which we could easily estimate. Supposing that the real concentration in normalized coordinates is $y_{\text{real}} = f(x)$, where f varies between 0 and 1, x_0 is computed by solving the equation $f(x_0) = t$, where t is a given threshold. The measured function, using immunofluorescence staining, is $y_{\text{meas}} = A_0 f(x) + B_0$, where B_0 is the background due to nonspecific attachment, and A_0 depends on the efficiency of probe–target interaction. The normalization procedure transforms $y_{\text{meas}}(x)$ into a normalized function $y_{\text{norm}}(x) = (y_{\text{meas}}(x) - B)/A$. x_{Bcd} is computed using $y_{\text{norm}}(x)$.

The error in the measurement of x_{Bcd} due to error in estimation of A and B is $\Delta x = (1/df/dx|_{x=x_0})(1-t)\Delta B/A_0$, where $df/dx|_{x=x_0}$ is the slope of $f(x)$ at the defining point, and ΔB is the absolute error in estimating B .

The key to reducing the error in the measurement of x_{Bcd} is to have a high signal-to-background ratio ($A/B \cong 8$ in our case). Using the above equation, we find $\Delta x_{\text{Bcd}} \cong 1\%$. The real variability of x_{Bcd} , Δ_n , is related to the measured one Δ_m and to error due to normalization Δ_{norm} through $[\Delta_m]^2 = \Delta_n^2 + \Delta_{\text{norm}}^2$. Hence, given the precision of our procedure, the error of normalization has a negligible effect on the measured variability of bcd gradient.

A more robust way of measuring the bcd variability is to use the exponential decay of Bcd concentration. Bcd profiles have a peak at about 35 μm (this peak is variable from embryo to embryo) and display a perfect exponential decay after twice this distance. Raw data of immunofluorescence for Bcd are fitted by $I = A \exp(-x/\lambda) + B$ for abscissae beyond twice the peak position. A nonlinear Levenberg–Marquardt fit procedure was used to estimate the parameters. The slope of the exponential, λ , for each embryo was then computed from all values of the raw curve, and did not depend on normalization parameters. The confidence limit of λ in each fitted curve was determined by the curvature matrix of the χ^2 function at its minimum¹⁷, and is smaller than 1 μm (σ_λ for between-embryo variability is 20 μm).

Received 28 August; accepted 4 December 2001.

1. Wolpert, L. Positional information and the spatial pattern of cellular differentiation. *J. Theor. Biol.* **25**, 1–47 (1969).
2. Driever, W. & Nusslein-Volhard, C. The bicoid protein determines position in the *Drosophila* embryo in a concentration-dependent manner. *Cell* **54**, 95–104 (1988).
3. Driever, W. & Nusslein-Volhard, C. A gradient of bicoid protein in *Drosophila* embryos. *Cell* **54**, 83–93 (1988).
4. Struhl, G., Struhl, K. & Macdonald, P. M. The gradient morphogen bicoid is a concentration-dependent transcriptional activator. *Cell* **57**, 1259–1273 (1989).
5. Lacalli, T. C. & Harrison, L. G. From gradient to segments: models for pattern formation in early *Drosophila*. *Semin. Dev. Biol.* **2**, 107–117 (1991).
6. Segel, I. H. *Enzyme Kinetics* (Wiley, New York, 1975).
7. Treisman, J. & Desplan, C. The products of the *Drosophila* gap genes *hunchback* and *Kruppel* bind to the hunchback promoters. *Nature* **341**, 335–337 (1989).
8. Wimmer, E. A., Carleton, A., Harjes, P., Turner, T. & Desplan, C. Bicoid-independent formation of thoracic segments in *Drosophila*. *Science* **287**, 2476–2479 (2000).
9. Struhl, G., Johnston, P. & Lawrence, P. A. Control of *Drosophila* body pattern by the hunchback morphogen gradient. *Cell* **69**, 237–249 (1992).
10. Jackle, H., Tautz, D., Schuh, R., Seifert, E. & Lehmann, R. Cross regulatory interactions among the gap genes of *Drosophila*. *Nature* **324**, 668–670 (1986).
11. Simpson-Brose, M., Treisman, J. & Desplan, C. Synergy between the hunchback and bicoid morphogens is required for anterior patterning in *Drosophila*. *Cell* **78**, 855–865 (1994).
12. Hulskamp, M., Lukowitz, W., Beermann, A., Glaser, G. & Tautz, D. Differential regulation of target genes by different alleles of the segmentation gene *hunchback* in *Drosophila*. *Genetics* **138**, 125–134 (1994).
13. St Johnston, D., Beuchle, D. & Nusslein-Volhard, C. *Staufen*, a gene required to localize maternal RNAs in the *Drosophila* egg. *Cell* **66**, 51–63 (1991).
14. Ferrandon, D., Elphick, L., Nusslein-Volhard, C. & St Johnston, D. *Staufen* protein associates with the 3'UTR of bicoid mRNA to form particles that move in a microtubule-dependent manner. *Cell* **79**, 1221–1232 (1994).
15. Roberts, D. B. (ed.) *Drosophila, A Practical Approach* (Oxford Univ. Press, Oxford, 1998).
16. Kossman, D., Small, S. & Reinitz, J. Rapid preparation of a panel of polyclonal antibodies to *Drosophila* segmentation proteins. *Dev. Genes Evol.* **208**, 290–294 (1998).
17. Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. *Numerical Recipes in C* (Cambridge Univ. Press, Cambridge, 1992).
18. Merrill, P., Sweeton, D. & Wieschaus, E. Requirements for autosomal gene activity during precellular stages of *Drosophila melanogaster*. *Development* **104**, 495–509 (1988).

Acknowledgements

Drosophila alleles were a gift from C. Desplan (FRT-*hb,nos*^{BN}), E. Gavis (*stau*^{D3}) and Nusslein-Volhard lab stock (*stau*⁹). This work has been partially supported by grants from the National Institutes of Health and the Howard Hughes Medical Institute. Discussions with C. Desplan, J. Grosshans, T. Lecuit, J. Reinitz and S. Small are here acknowledged.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.W. (e-mail: ewieschaus@molbio.princeton.edu).

Activation-induced cytidine deaminase turns on somatic hypermutation in hybridomas

Alberto Martin*, Philip D. Bardwell*, Caroline J. Woo*, Manxia Fan*, Marc J. Shulman† & Matthew D. Scharff*

* Department of Cell Biology, Albert Einstein College of Medicine, 1300 Morris Park Avenue Chanin 403, Bronx, New York 10461, USA

† Immunology Department, Medical Sciences Building, 1 King's College Circle, University of Toronto, Toronto, Ontario M5S 1A8, Canada

The production of high-affinity protective antibodies requires somatic hypermutation (SHM) of the antibody variable (V)-region genes. SHM is characterized by a high frequency of point mutations that occur only during the centroblast stage of B-cell differentiation. Activation-induced cytidine deaminase (AID), which is expressed specifically in germinal-centre centroblasts¹, is required for this process, but its exact role is unknown². Here we show that AID is required for SHM in the centroblast-like Ramos cells, and that expression of AID is sufficient to induce SHM in hybridoma cells, which represent a later stage of B-cell differentiation that does not normally undergo SHM. In one hybridoma, mutations were exclusively in G-C base pairs that were mostly within RGYW or WRCY motifs, suggesting that AID has primary responsibility for mutations at these nucleotides. The activation of SHM in hybridomas indicates that AID does not require other centroblast-specific cofactors to induce SHM, suggesting either that it functions alone or that the factors it requires are expressed at other stages of B-cell differentiation.

Three human B-cell lines (namely Ramos, BL-2 and CL-01) undergo SHM^{3–5}, thus opening the possibility of studying this process *in vitro*. We found previously⁶ that V-region mutation rates in different Ramos clones were correlated with the level of their AID messenger RNA, suggesting that AID is important in SHM in Ramos cells. Specifically, both the rates of mutation and the mRNA levels of AID for Ramos clones 6 and 7 were higher than those for Ramos clone 1 (Fig. 1a and ref. 6). To determine whether low AID expression was itself responsible for the low mutation rates in Ramos clone 1, this clone was stably transfected with either a vector expressing human AID (hAID) or an empty vector control. Mutation rates of typical transfected clones were then determined by sequencing unselected V regions after 1 or 2 months in culture (Table 1). Clones expressing low levels of AID (that is, clones C.1 and A.1) had very few mutations in the V region, whereas clones that expressed ~25-fold higher levels of AID mRNA (that is, clones A.2 and A.5) had many more V-region mutations (Fig. 1b and Table 1). Table 2 summarizes the mutational features of all the Ramos clones that expressed elevated levels of AID and shows that the rates and characteristics of the mutations in all of these clones were similar: there was a targeting bias of G/C nucleotides, transitions were slightly favoured over transversions, and ~35% of mutations were in G or C nucleotides within RGYW (A/G, G, C/T, A/T) or WRCY hot-spot sequences, motifs that are frequently targeted in SHM both *in vivo* and *in vitro*^{7,8}. These results indicate that AID is required for

SHM in Ramos cells.

Ramos cells express surface markers that suggest that their normal cellular counterpart is a germinal-centre centroblast³, which is the cell type that normally undergoes SHM. Although AID is required for SHM in these cells, other factors specific to centroblasts might also be required. To test this notion, we determined whether AID could induce SHM in hybridomas, which represent plasma cells that are beyond the developmental stage that performs SHM. We first examined the N89 and N114 hybridomas because they have nonsense codons within the V region of their endogenous antibody heavy-chain gene (top of Fig. 2a and ref. 9), allowing us to assay many independently transfected clones by assaying for nonsense-codon revertants by using the ELISA-spot assay⁶.

N89 and N114 cells were stably transfected with the human AID (hAID) expression vector; individual drug-resistant colonies were expanded and assayed for secreted immunoglobulin- μ (IgM) with the ELISA-spot assay. Each ELISA spot indicated that a cell had reverted the nonsense codon and was secreting antibody (Fig. 2a inset and ref. 10). Figure 2a shows the frequency of revertants identified for each individual clone. None of the N89 and N114 clones that were transfected with the empty vector displayed a revertant frequency above 10^{-6} (Fig. 2a). However, more than 50% of individual N89 and N114 clones transfected with the hAID construct had revertant frequencies higher than 10^{-6} (Fig. 2a).

To determine the rate of mutation at the nonsense codon of N114, 12 subclones of the AID-positive N114 A.3 clone were analysed for revertants (Fig. 2a, right panel) yielding a mutation rate of 1.4×10^{-6} mutations per base pair per generation, as calculated by fluctuation analysis⁶. As shown below, this reversion rate greatly underestimates the mutation rate for the V region as a whole for two reasons: (1) the nonsense codon for N114 (and for N89) cells is not within an RGYW or WRCY hot-spot motif, and (2) most mutations observed in these hybridoma clones were transition mutations in G/C nucleotides (see below), which would convert the TGA nonsense codon to TAA, another nonsense codon.

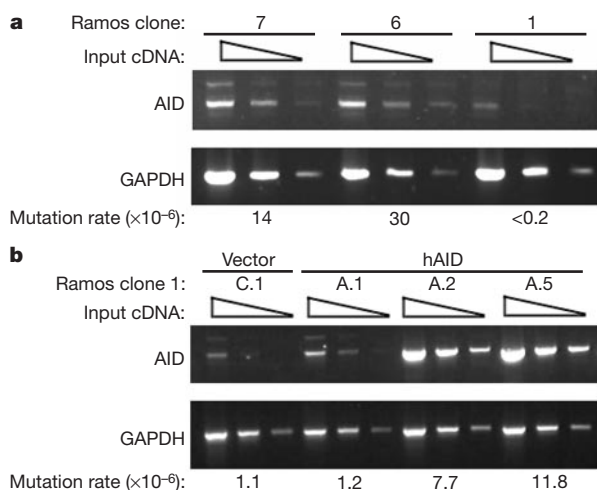


Figure 1 hAID expression induces SHM in non-mutating Ramos cells. RNA was isolated from selected subclones of Ramos (a) and stable transfectants of the non-mutating Ramos clone 1 that overexpress hAID (b). hAID and GAPDH were reverse-transcribed, and fivefold dilutions of the resulting cDNA were amplified by PCR (see Methods). Levels of AID mRNA in clones 6 and 7 were about fivefold higher than in clone 1 (ref. 6) (a) and ~25-fold higher in clones A.2 and A.5 than in C.1 and A.1 (b). The mutation rates shown for the subclones in a are taken from ref. 6, and those in b are from the present study.

Some N89 and N114 clones transfected with the hAID expression vector did not revert at a detectable frequency (Fig. 2a). Northern blots revealed that all tested clones that did not express AID did not revert above 10^{-6} (Fig. 2b). However, some clones that expressed AID (namely N89 A.3, A.5, A.10 and A.16, and N114 A.2 and A.8; Fig. 2b) also did not revert above 10^{-6} . Although this suggests that AID does not induce SHM in all hybridoma clones, this is probably due to the inherently stochastic nature of this analysis: a nonsense revertant that arises early in the propagation of a clone will result in a culture that has accumulated many revertants, whereas a revertant that arises late during the propagation of the clone will be represented by very few revertants¹¹. This is exemplified by the high range of revertant frequencies in subclones of N114 A.3, with

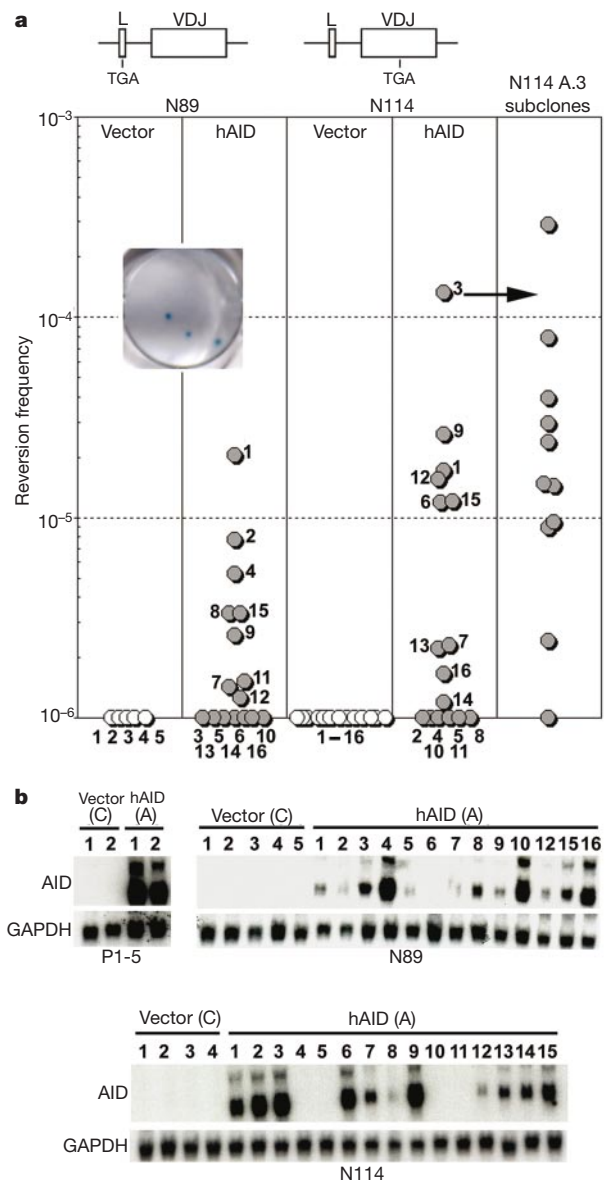


Figure 2 AID induces SHM in hybridomas. **a**, Hybridomas N89 (nonsense in leader) and N114 (nonsense in V region) were transfected with empty vector or the hAID construct. The frequency of nonsense revertants, as detected with the ELISA-spot assay, is plotted. Typical ELISA spots for N89 are shown in the inset. The difference in the frequencies of reversion of N114 vector and N114 hAID was statistically significant ($P < 0.05$). **b**, Northern blots for hAID and GAPDH of N89, N114 and P1-5 transfected hybridoma clones. Clone numbers for N89 and N114 clones correspond to numbers in a.

Clone	Culture duration (months)	V-region mutations*	Total base pairs sequenced	10 ⁴ × Frequency (mutations per base pair)	Mutated V regions/total	10 ⁶ × Mutation rate† (mutations per base pair per generation)
Ramos C.1	1	0	11,200	<0.89	0/26	<2.5
Ramos C.1	2	1	12,900	0.78	1/30	1.1
Ramos A.1‡	2	1	11,600	0.86	1/27	1.2
Ramos A.2	1	5	12,900	3.9	4/30	10.7
Ramos A.2	2	7	12,500	5.6	7/29	7.7
Ramos A.5	1	6	16,300	3.7	4/38	10.2
Ramos A.5	2	13	15,300	8.5	9/31	11.8
P1-5 C.1	2	1	11,900	0.84	1/35	1.4
P1-5 C.2	2	0	7,140	<1.4	0/21	<2.3
P1-5 A.1	2	28	28,900	9.3	22/85	15.7
P1-5 A.2	2	6	5,780	10.4	6/17	17.3
N114 C.1	1	0	10,980	<0.91	0/18	<3.0
N114 A.3	1	12	21,960	5.5	11/36	18.2

‡Expression of AID was low in this clone.

To measure the real rate of mutation in these AID-expressing hybridomas, transfected clones were grown for 2 months and

These findings show that AID is sufficient to activate SHM in plasma-like cells, suggesting that its activity does not depend on other centroblast-specific factors. We have previously used plasma-like NSO cells to measure SHM of immunoglobulin transgenes that contain V-region nonsense codons¹⁰. Sequencing unselected V regions revealed no mutations of the transgenes in 2-month-old cultures, indicating low mutation rates in NSO cells. However, fluctuation analysis for nonsense revertants revealed a wide distribution of mutation rates¹⁰ that depended strongly on the position of the nonsense codon in the V region, its proximity to hot-spot motifs and the number of transgenes present in the transfectant. In particular we did not detect expression of AID in NSO cells (data not shown). Although we cannot rule out the transient expression of AID in NSO, it is possible that the high rate of reversion that we observed in a few NSO transfectants was engendered by the insertion site and/or the presence of the transgene in a repetitive structure. Nevertheless, we estimate that the overall mutation rates in NSO are low compared with those in Ramos clones and

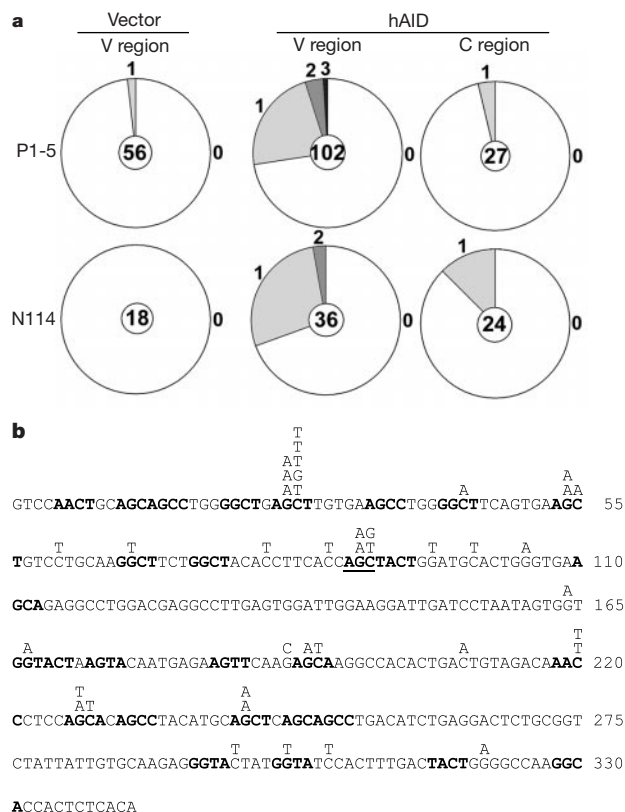


Figure 3 Mutations observed in hybridoma clones. **a**, Pie charts, as shown previously³, depicting the distribution of the frequencies of mutation of P1-5 and N114 hybridoma clones. Shown are the number of sequences analysed (centre of pie) and the proportion of sequences with 0, 1, 2 ... mutations (pie slices). **b**, All mutations located within the V region (V186.2) of hybridoma P1-5 clones. Duplicate mutations were counted once in Table 1 unless genealogies indicated that mutations were unique. Hot-spot motifs (RGYW and WRCY) are shown in bold. Although other hot-spot motifs are frequently mutated, we and others have observed a high frequency of mutation at the codon 31 hot spot (underlined) *in vivo*²⁷.

Table 2 Characteristics of mutations observed in Ramos clones 6 and 7 and in hAID-transfected cell lines

	Ramos clones 6 and 7*	Ramos clones A.2 and A.5	P1-5 clones A.1 and A.2	N114 clone A.3
GC mutations/total	40/51 (78%)	25/31 (81%)	34/34 (100%)	9/12 (75%)
Transitions†/total	23/51 (49%)	13/31 (42%)	24/34 (71%)	7/12 (58%)
Deletions/total	2/51 (4%)	2/31 (6%)	0/34 (0%)	0/12 (0%)
Hot spots‡/total	18/51 (35%)	10/31 (32%)	21/34 (62%)§	6/12 (50%)

* Previously published sequences⁶.

† C → T or G → A.

‡ Mutations at G or C nucleotides within RGYW or WRCY are defined as hot-spot mutations; 7% (36/550), 9% (32/340) and 6% (39/610) of total nucleotides in the V regions of Ramos, P1-5 and N114 clones, respectively, were hot-spot nucleotides.

§ Statistically significant when compared with Ramos clones A.2 and A.5 ($P < 0.05$ for GC mutations, $P < 0.001$ for transitions and for hot-spot mutations).

hybridomas expressing AID. Taken together with recent findings that the pre-B-cell line 18.81 expresses AID and performs SHM¹⁷, AID might be able to function at any developmental B-cell stage, or perhaps even in non-B cells. This implies that AID functions alone or with cofactors that are expressed ubiquitously. In fact, many of the factors that have been shown to be involved in SHM, such as MutS homologue 2 (MSH2)^{15,18–20}, MSH6 (ref. 21) and DNA polymerases ζ ²² and η ^{23,24} are enzymes expressed in most cells.

The findings reported here also have practical implications. The fact that hybridomas can be induced to undergo high rates of SHM with expression of AID might allow one to obtain subclones that produce high-affinity monoclonal antibodies and/or antibodies that are more specific. This has been a goal that has been sought by many²⁵ since Kohler and Milstein first described the hybridoma technology²⁶. □

Methods

Cell lines, cell culture and transfection conditions

Ramos cells were grown as described previously⁶. N89 and N114 hybridoma cells were described previously⁹; the hybridoma P1-5 was obtained from A. Bothwell¹⁴. Ramos cells were electroporated with 10 μ g linearized DNA in IMDM (Iscove's modified Dulbecco's medium) at 250 V and 960 μ F, plated into 96-well plates, and selected with 0.6 mg ml^{−1} hygromycin B. The hybridoma cells were electroporated as described previously¹⁰ and selected in 0.3 mg ml^{−1} hygromycin B. The ELISA-spot assay was performed as reported previously⁶. In brief, each drug-resistant colony was expanded to $\sim(1-5) \times 10^6$ cells and plated on 96-well plates precoated with anti-mouse IgM antibody. After 22 h the plates were developed for secreted IgM.

Constructs

Full-length hAID was amplified by using primers reported previously⁶ and cloned into Zero BluntTOPO polymerase chain reaction (PCR) Cloning Kit (Invitrogen) to sequence. The hAID insert was excised with *Eco*R1, blunted with Klenow polymerase and cloned into pCEP4 (Invitrogen) digested with *Pvu*II. Vectors were digested with *Nru*I and *Eco*RV before transfection.

PCR amplification, cloning and sequencing

Genomic DNA was prepared as reported previously⁶. V regions from the various B-cell lines were amplified with *Pfu* polymerase (Stratagene) from genomic DNA by using 30 cycles of 95 °C for 15 s, 56 °C for 15 s and 72 °C for 30 s. Primers for the N114 V region were TTACCTGGGTCTATGGCAGT (5' primer) and TGAAGGCTCAGAATCCCCC (3' primer), and for the C μ 2-3 region were CCCCTCCTTTGCCGACATCTTCC (5' primer) and TTCCATTCTCCTCGTCACAGTC (3' primer). Primers for Ramos and the P1-5 V region have been published previously^{6,27}. Primers for the P1-5 C γ 1 exons 2 and 3 were CACACAGCTCAGACGCAACCCC (5' primer) and GGATCATTTACCGAGAGAGTGG-GAGAGG (3' primer). This primer pair amplifies both the Balb/c and the C57Bl/6 C γ 1 segments from the P1-5 hybridoma, which can be distinguished by allelotypic differences. Because the nitrophenyl-specific V region of P1-5 is derived from C57Bl/6 (ref. 14), only C-region sequences from C57Bl/6 are shown in Fig. 3a. PCR products were cloned and sequenced as described previously⁶.

Extraction of RNA, PCR and northern blots

About 5×10^6 cells were lysed with 1 ml Trizol reagent (Gibco BRL) and RNA was extracted in accordance with the manufacturer's instructions. About 1 μ g total RNA was either run on formaldehyde gels for northern blots or reverse transcribed with the Superscript II kit (Gibco BRL). The reverse-transcribed product (5 μ l) was diluted fivefold with water sequentially three times; 1 μ l of each of these three dilutions was used in a PCR reaction, and all amplifications of each cDNA from each different clone were done together. *Taq* polymerase (Roche) was used to amplify glyceraldehyde phosphate dehydrogenase (GAPDH) and AID by using primers and conditions described previously⁶.

Statistics

Statistics for sequencing data in Table 2 and primary data for reversion frequencies in Fig. 2a were measured by the independent-samples *t*-test with equal variances assumed (SPSS version 10).

Received 11 December 2001; accepted 9 January 2002.

Published online 30 January 2002, DOI 10.1038/nature714.

- Muramatsu, M. *et al.* Specific expression of activation-induced cytidine deaminase (AID), a novel member of the RNA-editing deaminase family in germinal centre B cells. *J. Biol. Chem.* **274**, 18470–18476 (1999).
- Kinoshita, K. & Honjo, T. Linking class-switch recombination with somatic hypermutation. *Nature Rev. Mol. Cell Biol.* **2**, 493–503 (2001).
- Sale, J. E. & Neuberger, M. S. TdT-accessible breaks are scattered over the immunoglobulin V domain in a constitutively hypermutating B cell line. *Immunity* **9**, 859–869 (1998).
- Denépoux, S. *et al.* Induction of somatic mutation in a human B cell line in vitro. *Immunity* **6**, 35–46 (1997).
- Zan, H. *et al.* Induction of Ig somatic hypermutation and class switching in a human monoclonal IgM⁺ IgD⁺ B cell line in vitro: definition of the requirements and modalities of hypermutation.. *J. Immunol.* **162**, 3437–3447 (1999).
- Zhang, W. *et al.* Clonal instability of V region hypermutation in the Ramos Burkitt's lymphoma cell line. *Int. Immunol.* **13**, 1175–1184 (2001).
- Wagner, S. D., Milstein, C. & Neuberger, M. S. Codon bias targets mutation. *Nature* **376**, 732 (1995).
- Rogozin, I. B. & Kolchanov, N. A. Somatic hypermutagenesis in immunoglobulin genes. II. Influence of neighbouring base sequences on mutagenesis. *Biochim. Biophys. Acta* **1171**, 11–18 (1992).
- Connor, A., Wiersma, E. & Shulman, M. J. On the linkage between RNA processing and RNA translatability. *J. Biol. Chem.* **269**, 25178–25184 (1994).
- Lin, M. M., Zhu, M. & Scharff, M. D. Sequence dependent hypermutation of the immunoglobulin heavy chain in cultured B cells. *Proc. Natl Acad. Sci. USA* **94**, 5284–5289 (1997).
- Luria, S. E. & Delbruck, M. Mutation of bacteria from virus sensitivity to virus resistance. *Genetics* **28**, 491–511 (1943).
- Rada, C. & Milstein, C. The intrinsic hypermutability of antibody heavy and light chain genes decays exponentially. *EMBO J.* **20**, 4570–4576 (2001).
- Rada, C., Yelamos, J., Dean, W. & Milstein, C. The 5' hypermutation boundary of κ chains is independent of local and neighbouring sequences and related to the distance from the initiation of transcription. *Eur. J. Immunol.* **27**, 3115–3120 (1997).
- Tao, W. & Bothwell, A. L. Development of B cell lineages during a primary anti-hapten immune response. *J. Immunol.* **145**, 3216–3222 (1990).
- Rada, C., Ehrenstein, M. R., Neuberger, M. S. & Milstein, C. Hot spot focusing of somatic hypermutation in MSH2-deficient mice suggests two stages of mutational targeting. *Immunity* **9**, 135–141 (1998).
- Spencer, J., Dunn, M. & Dunn-Walters, D. K. Characteristics of sequences around individual nucleotide substitutions in IgVH genes suggest different GC and AT mutators. *J. Immunol.* **162**, 6596–6601 (1999).
- Bachl, J., Carlson, C., Gray-Schopfer, V., Dessing, M. & Olsson, C. Increased transcription levels induce higher mutation rates in a hypermutating cell line. *J. Immunol.* **166**, 5051–5017 (2001).
- Phung, Q. H. *et al.* Increased hypermutation at G and C nucleotides in immunoglobulin variable genes from mice deficient in the MSH2 mismatch repair protein. *J. Exp. Med.* **187**, 1745–1751 (1998).
- Reynaud, C. A. *et al.* Mismatch repair and immunoglobulin gene hypermutation: did we learn something? *Immunol. Today* **20**, 522–527 (1999).
- Vora, K. A. *et al.* Severe attenuation of the B cell immune response in Msh2-deficient mice. *J. Exp. Med.* **189**, 471–481 (1999).
- Wiesendanger, M., Kneitz, B., Edelmann, W. & Scharff, M. D. Somatic mutation in MSH3, MSH6, and MSH3/MSH6-deficient mice reveals a role for the MSH2–MSH6 heterodimer in modulating the base substitution pattern. *J. Exp. Med.* **191**, 579–584 (2000).
- Zan, H. *et al.* The translesional polymerase ζ plays a major role in Ig and Bcl-6 somatic mutation. *Immunity* **14**, 643–653 (2001).
- Zeng, X. *et al.* DNA polymerase η is an A-T mutator in somatic hypermutation of immunoglobulin variable genes. *Nature Immunol.* **2**, 537–541 (2001).
- Rogozin, I. B., Pavlov, Y. I., Bebenek, K., Matsuda, T. & Kunkel, T. A. Somatic mutation hotspots correlate with DNA polymerase η error spectrum. *Nature Immunol.* **2**, 530–536 (2001).
- Kobrin, B. J., Casadevall, A., Pirofski, L. A., Schiff, C. & Scharff, M. D. In *Somatic Hypermutation in V Regions* (ed. Steele, E. J.) 11–28 (CRC, Boca Raton, 1990).
- Kohler, G. & Milstein, C. Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature* **256**, 495–497 (1975).
- Sack, S. Z., Bardwell, P. D. & Scharff, M. D. Testing the reverse transcriptase model of somatic mutation. *Mol. Immunol.* **38**, 303–311 (2001).

Acknowledgements

We thank B. Diamond, J. Warner, M. Goodman and R. Laskov for critical review of the manuscript, and A. Bothwell for providing the P1-5 hybridoma. This work was supported by grants from the National Institutes of Health to P.D.B., to C.J.W. and to M.D.S., who is also supported by the Harry Eagle chair provided by the National Women's Division of the Albert Einstein College of Medicine. A.M. is a recipient of Cancer Research Institute and Harry Eagle fellowships.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.D.S. (e-mail: scharff@aecom.yu.edu). Accession numbers for mutated sequences of Ramos clones 6 and 7 are AF385858–AF385893, and those of Ramos clones A.2 and A.5, P1-5 clone A.1 and N114 clone A.3 are AF439565–AF439624, and those of P1-5 clone A.2 are AF455739–AF455745.

Lateral relocation of auxin efflux regulator PIN3 mediates tropism in *Arabidopsis*

Jiří Friml^{†‡}, Justyna Wiśniewska^{‡§}, Eva Benková^{*}, Kurt Mendgen[§] & Klaus Palme^{*||}

^{*} Max-Delbrück-Laboratorium in der Max-Planck-Gesellschaft, 50829 Köln, Germany

[†] Zentrum für Molekularbiologie der Pflanzen, Universität Tübingen, 72076 Tübingen, Germany

[‡] Department of Biotechnology, Institute of General and Molecular Biology, 87–100 Toruń, Poland

[§] Lehrstuhl Phytopathologie, Universität Konstanz, 78457 Konstanz, Germany

^{||} Institut für Biologie II, Universität Freiburg, 79104 Freiburg, Germany

Long-standing models propose that plant growth responses to light or gravity are mediated by asymmetric distribution of the phytohormone auxin^{1–3}. Physiological studies implicated a specific transport system that relocates auxin laterally, thereby effecting differential growth⁴; however, neither the molecular components of this system nor the cellular mechanism of auxin redistribution on light or gravity perception have been identified. Here, we show that auxin accumulates asymmetrically during differential growth in an efflux-dependent manner. Mutations in the *Arabidopsis* gene *PIN3*, a regulator of auxin efflux, alter differential growth. *PIN3* is expressed in gravity-sensing tissues, with *PIN3* protein accumulating predominantly at the lateral cell surface. *PIN3* localizes to the plasma membrane and to vesicles that cycle in an actin-dependent manner. In the root columella, *PIN3* is positioned symmetrically at the plasma membrane but rapidly relocates laterally on gravity stimulation. Our data indicate that *PIN3* is a component of the lateral auxin transport system regulating tropic growth. In addition, actin-dependent relocalization of *PIN3* in response to gravity provides a mechanism for redirecting auxin flux to trigger asymmetric growth.

Plants orientate their growth with respect to the direction of light (phototropism) or gravity (gravitropism)¹. As early as 1926 a widely accepted model for plant tropisms, the Cholodny–Went hypothesis, was presented². It proposes differential distribution of the plant hormone auxin in lateral direction on gravity or light stimulation. Subsequently, different auxin levels elicit differential growth rates, which ultimately lead to bending of the shoot or root³. Visualization of asymmetrically distributed auxin response in gravistimulated tobacco stems⁵ and *Arabidopsis* roots⁶ experimentally supported

this hypothesis. Polar auxin transport represent a plausible means of lateral auxin distribution, as its chemical inhibition affects differential growth responses such as tropisms and apical hook formation^{7,8}. Physiologically characterized components of polar auxin transport are cellular efflux carriers, whose polar localization within cells is thought to determine the direction of auxin flux⁹. The recently identified *PIN* genes of *Arabidopsis* appear to encode essential components of these carriers⁷. A role of *PIN2* in regulation of basipetal auxin transport and gravitropism in root^{6,10,11} as well as a role of *PIN1* in basipetal auxin transport in the stem have been reported¹²; however, so far the molecular basis of shoot tropic responses remains elusive. Lateral auxin transport with a specific, laterally localized auxin efflux carrier was proposed⁴ to explain the exchange of auxin between vasculature, where the main basipetal auxin stream occurs¹³, and peripheral tissues controlling elongation¹⁴. Nevertheless the lack of any molecular data supporting this concept still leaves the existence of such a system in question.

We analysed the relationship between auxin efflux, auxin redistribution and differential growth responses in *Arabidopsis* hypocotyls. Growth responses were examined and auxin levels indirectly visualized using the synthetic *DR5::GUS* auxin reporter¹⁵, whose activity correlates with direct auxin measurements^{16,17}. Seedlings grown under standard conditions displayed normal phototropic and gravitropic curvature accompanied by an asymmetric *DR5::GUS* expression suggesting elevated auxin levels on the more elongated side of the hypocotyl (Fig. 1a, c). In contrast, seedlings that did not undergo light or gravity stimulation showed lower and uniform expression of the *DR5::GUS* reporter (data not shown). Seedlings subjected to light and gravity stimulation with simultaneous auxin efflux inhibition failed to show any differential *DR5::GUS* expression as well as tropic curvature (Fig. 1b, d). Differential *DR5::GUS* expression was also detected during the process of apical hook formation (Fig. 1e), and seedlings grown on auxin efflux inhibitors failed to show asymmetric auxin distribution and apical hook formation (data not shown). These results show that asymmetric growth of the shoot correlates with auxin efflux-dependent asymmetric auxin distribution, and implicate the existence of auxin efflux components involved in asymmetric growth responses.

The auxin efflux carrier candidates are *PIN* proteins, although their function in the regulation of differential growth has not been reported^{7,8}. We screened for new members of the *PIN* gene family and analysed corresponding knockout mutants for differential growth defects. One of the genes, *PIN3*, was isolated from genomic and complementary DNA libraries with probes derived from the conserved region of *PIN1*. The deduced *PIN3* protein shares 67%

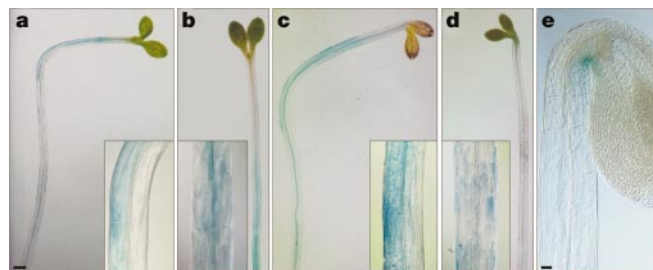


Figure 1 Auxin response during differential hypocotyl growth. **a–d**, Expression of the *DR5::GUS* reporter in hypocotyl of untreated (**a, c**) or auxin efflux inhibitor (NPA)-treated (**b, d**) wild-type seedlings upon stimulation by light (**a, b**) or gravity (**c, d**). Insets show details of *DR5::GUS* expression. Scale bars, 400 μ m. **e**, Asymmetric expression of *DR5::GUS* reporter in the apical hook of an etiolated seedling. Scale bar, 50 μ m.

identity to PIN1, displays the characteristic three-domain topology with about 10 transmembrane domains characteristic of all PIN proteins^{7,8}, and shows similarity to bacterial transporters and detoxification carriers, suggesting the same biochemical function for PIN proteins in auxin efflux^{7,8}. By a reverse genetic strategy¹⁸ two mutant alleles designated as *pin3-1* and *pin3-2* were identified (Fig. 2a). From the progeny of *pin3-2* a stable, footprint mutant allele exhibiting a deletion of two base pairs, designated *pin3-3*, was isolated. *PIN3* expression (data not shown) and immunolocalization (Fig. 4e) studies demonstrated that all three alleles were null.

pin3 mutants display defects in differential growth. Gravitropic and phototropic responses as well as apical hook maintenance are reduced in *pin3* mutants (Fig. 2b–d and Table 1). *pin3* seedlings also display shorter hypocotyls and roots than wild-type seedlings when grown in light (Fig. 2f and Table 1). However, no such differences

were observed in dark-grown seedlings (Fig. 2e and Table 1). Shorter hypocotyl epidermis cells in *pin3* suggested that this phenotype was caused by a defect in cell elongation (Fig. 2g, h and Table 1). The specificity of the *pin3* phenotype was further confirmed by identification of two revertants displaying wild-type phenotype among homozygous *pin3-2* seedlings. Sequencing of the *PIN3* locus in both plants confirmed the excision of the *En-1* transposon and restoration of the wild-type *PIN3* coding sequence. All defects in *pin3* mutants can be mimicked by growing wild-type plants on auxin efflux inhibitors (Fig. 2i and Table 1)^{7,19,20}, further supporting a role for PIN3 in auxin efflux. The *pin3* mutation does not abolish differential growth to the same extent that auxin efflux inhibitors at increased concentrations do. This suggests some level of functional redundancy between PIN3 and other PIN protein(s).

Northern blot analysis revealed the presence of *PIN3* transcripts

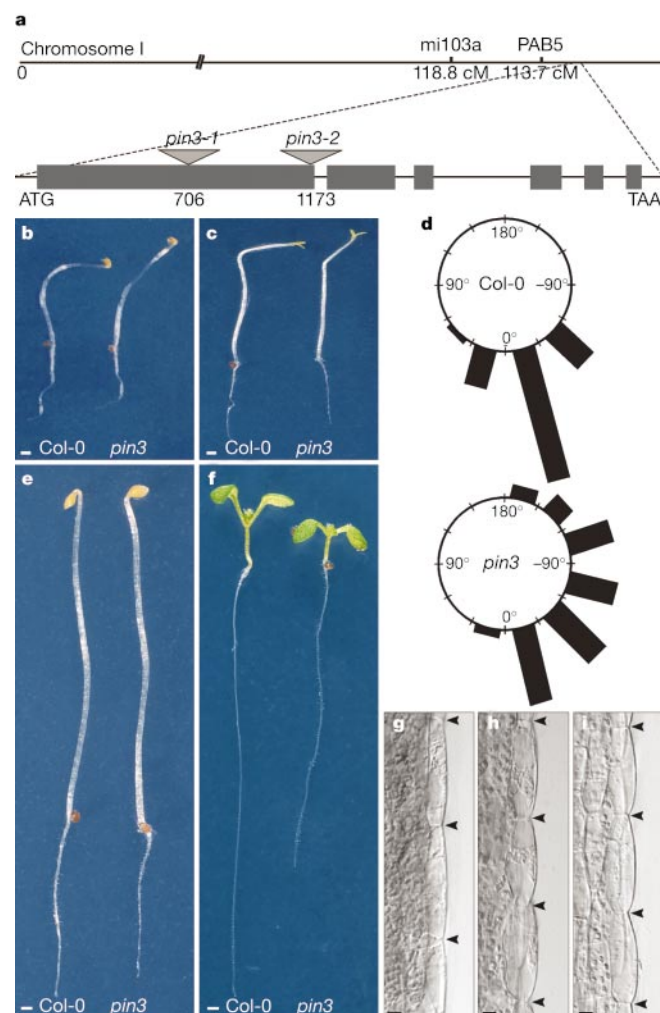


Figure 2 *pin3* mutants. **a**, Exon/intron organization of *PIN3* on chromosome 1 showing positions of *En-1* elements in *pin3-1* and *-2* alleles. **b, c**, *pin3* hypocotyls are defective in gravitropic (**b**) as well as phototropic (**c**) responses. Scale bars, 500 μ m. **d**, *pin3* mutants are defective in root gravitropism. Each gravistimulated root was assigned to one of twelve 30° sectors. The length of each bar represents the percentage of seedlings showing direction of root growth within that sector. **e, f**, Dark-grown *pin3* seedlings (**e**) have the same length of hypocotyls and roots as wild type, but display defects in apical hook maintenance. Light-grown *pin3* seedlings (**f**) display shorter hypocotyls and roots compared with wild type. Scale bars, 500 μ m. **g–i**, Hypocotyl epidermis cells of *pin3* (**h**) and NPA-grown wild-type (**i**) seedlings show reduced length compared with untreated wild type (**g**). Arrowheads indicate cell boundaries. Scale bars, 10 μ m. The following *pin3* alleles are depicted: *pin3-1* (**b**), *pin3-2* (**c, d**) and *pin3-3* (**e, f, h**).

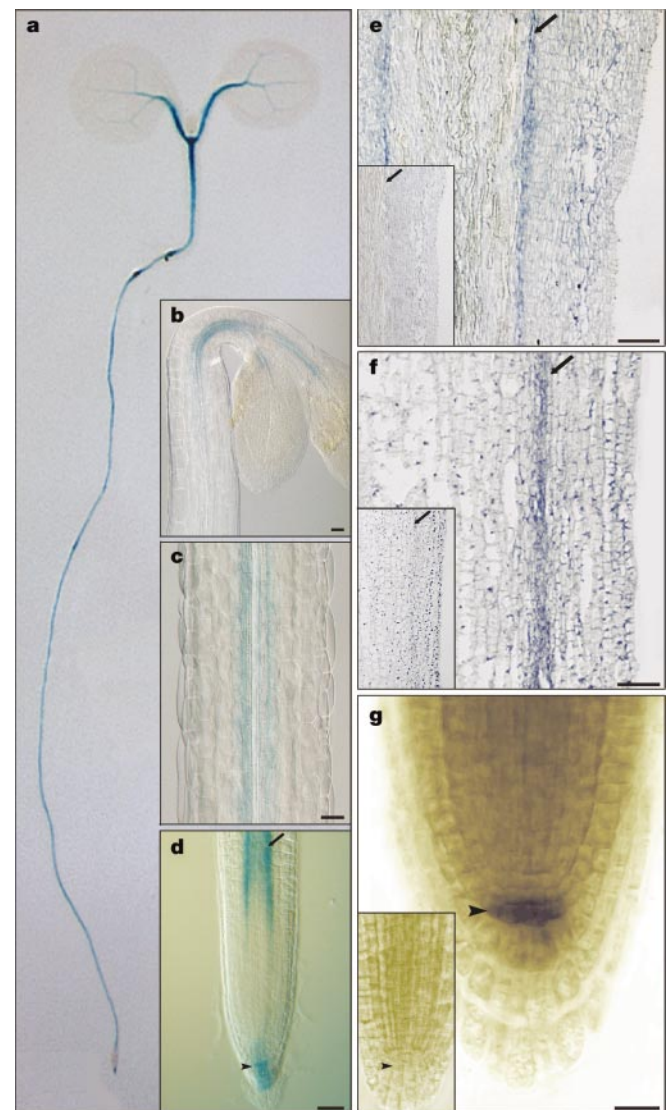


Figure 3 Expression of the *PIN3* gene. **a–d**, *PIN3::GUS* transgenic seedlings show *GUS* staining in the apical hook (**b**), around the hypocotyl vasculature (**c**), and in the root pericycle (arrow) and columella (arrowhead) (**d**). Scale bars, 50 μ m. **e–g**, *PIN3* mRNA was detected by *in situ* hybridization in hypocotyl (**e**), stem (**f**) starch sheath (arrows) and root columella (arrowheads) (**g**) cells. Insets show sense controls (**e–g**). Scale bars: 200 μ m (**e**), 100 μ m (**f**) and 20 μ m (**g**). Longitudinal sections (**e, f**) and whole-mount preparations (**a–d, g**) are shown.

Table 1 Quantitative effects of *pin3* mutations and auxin efflux inhibition

Measurement	Col-0	(n)	<i>pin3</i>	(n)	NPA*	(n)
Hypocotyl length in light (mm)	2.21 ± 0.16	(58)	1.68 ± 0.15	(66)	1.43 ± 0.14	(56)
Root length in light (cm)	2.41 ± 0.22	(58)	1.70 ± 0.16	(66)	1.67 ± 0.16	(56)
Hypocotyl length in dark (cm)	1.32 ± 0.18	(54)	1.35 ± 0.12	(62)	1.41 ± 0.15	(62)
Root length in dark (cm)	0.77 ± 0.07	(54)	0.78 ± 0.14	(62)	0.85 ± 0.11	(62)
Hypocotyl cell length (μm)	101.8 ± 10.2	(63)	75.5 ± 8.40	(54)	72.2 ± 7.90	(61)
Hypocotyl gravitropism (deg)	78.1 ± 16.9	(78)	31.8 ± 15.0	(81)	—	—
Hypocotyl phototropism (deg)	72.5 ± 15.5	(65)	40.0 ± 9.70	(72)	—	—
Opened apical hooks (%)	16.2 ± 9.67	(81)	65.4 ± 8.02	(97)	—	—

Wild-type plants treated with NPA.

predominantly in the shoot and root (data not shown). The staining of transgenic seedlings carrying the *PIN3* promoter fused to the β -glucuronidase (*GUS*) gene revealed *PIN3::GUS* expression associated with vasculature (Fig. 3a). Detailed inspection revealed *GUS* staining in the apical hook of the etiolated seedling (Fig. 3b), in the shoot endodermis (starch sheath) around the vasculature (Fig. 3c), and in the root pericycle and columella (Fig. 3d). To localize the *PIN3* protein, *PIN3*-specific antibodies were raised and immunolocalization experiments performed. The *PIN3* protein was found at the periphery of shoot starch sheath as well as root pericycle and columella cells (Fig. 4a–e). Longitudinal sections revealed that *PIN3* is localized in a polar way predominantly at the lateral, inner side of the starch sheath as well as root pericycle cells (Fig.

4c, d). In a minor portion of the cells an additional basal or uniform localization of *PIN3* was observed (data not shown). In contrast, only uniform distribution of *PIN3* in root columella cells was detected (Fig. 4e). No *PIN3* signal was detected in *pin3* knockout mutants (Fig. 4e, inset). *PIN3* messenger RNA localization using *in situ* hybridization confirms the *GUS* and immunolocalization staining patterns (Fig. 3e–g).

Anti-*PIN3* immunogold labelling revealed signals at the plasma membrane (Fig. 5a) and frequently in vesicles of about 70 nm diameter, similar in size to exocytotic vesicles (Fig. 5b), suggesting a dynamic subcellular movement of *PIN3*. The *PIN3* protein appears to cycle rapidly between the plasma membrane and undefined endosomal compartments, since the incubation of seedlings with the exocytosis inhibitor Brefeldin A (BFA) led to internalization of *PIN3* (Fig. 5d, compare with c), also in the presence of the protein synthesis inhibitor cycloheximide (Fig. 5f, compare with e). Pericycle cells show an accumulation of the *PIN3* label in perinuclear compartments (Fig. 5d), very similar to previously described BFA compartments²¹. In contrast, in columella cells *PIN3* internalizes in smaller compartments without any regular

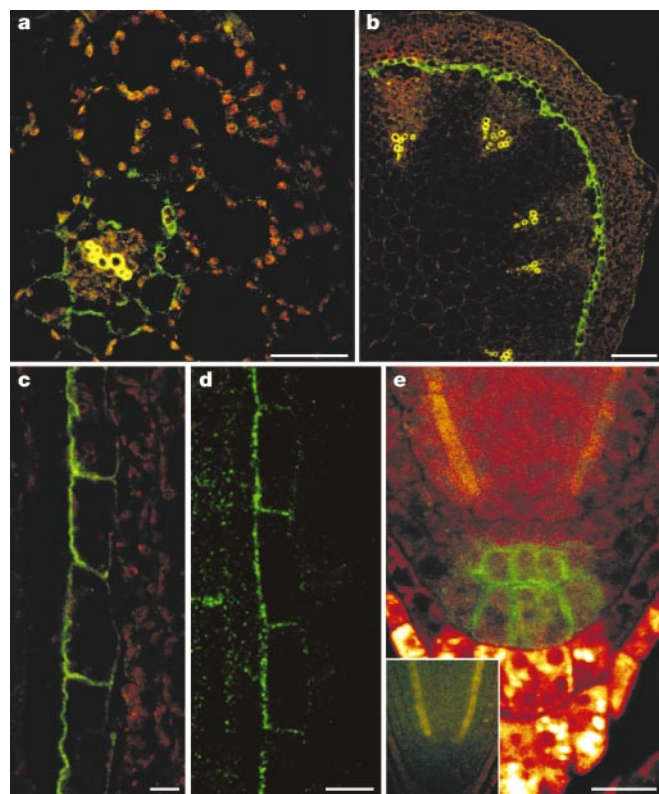


Figure 4 Localization of the *PIN3* protein. **a, b**, *PIN3* protein localizes to the surface of hypocotyl (**a**) and stem (**b**) starch sheath cells. Scale bars, 50 μm (**a**) and 100 μm (**b**). **c, d**, Longitudinal view of stem starch sheath (**c**) and root pericycle (**d**) cells shows lateral localization of *PIN3* at the inner cell surface. Scale bars: 10 μm (**c**) and 5 μm (**d**). **e**, In the root tip, *PIN3* is found uniformly in the columella cell boundaries. Inset shows no *PIN3* staining in the *pin3* mutant. Scale bar, 20 μm. Transversal (**a, b**) and longitudinal (**c**) sections, and whole-mount (**d, e**) preparations are shown. The *PIN3* protein signal is visualized by green coloration.

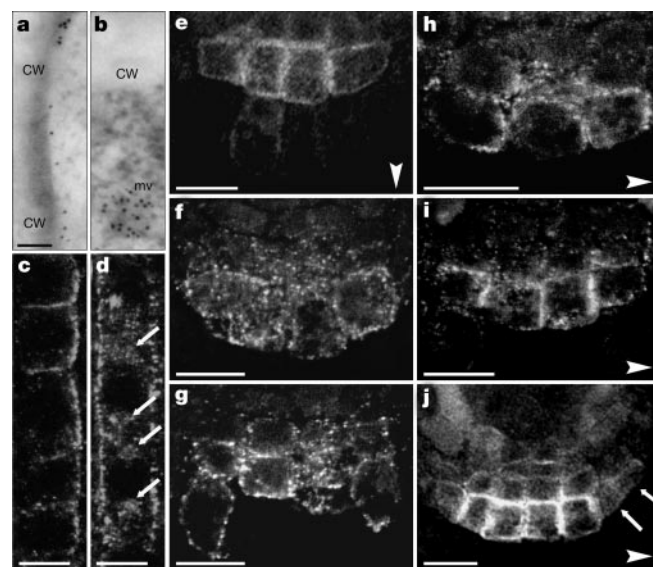


Figure 5 Subcellular localization, actin-dependent cycling and gravity-dependent asymmetric relocation of the *PIN3* protein. **a, b**, Immunogold electron microscopy shows *PIN3* at the plasma membrane (**a**) and in vesicles (**b**). cw, cell wall; mv, membrane vesicle. Scale bars, 50 nm. **c–g**, Actin-dependent cycling of *PIN3*. Localization of *PIN3* in columella (**e–g**) and pericycle (**c, d**) cells. Internalization of *PIN3* after BFA (**d**), cycloheximide/BFA (**f**) and latrunculin B (**g**) treatments. Untreated controls (**c, e**) are shown. Perinuclear compartments are indicated by arrows in **d**. Scale bars, 5 μm (**c, d**) and 10 μm (**e–g**). **h–j**, Relocation of *PIN3* in columella cells after a change in the gravity vector after 2 min (**h**) and 10 min (**i**). After 1 h the *PIN3* localization domain expands in lateral root cap (indicated by arrows) and in columella initials (**j**). Scale bars, 10 μm. The apparent gravity vectors are indicated by arrowheads.

positioning (Fig. 5f). A similar pattern of PIN3 localization was observed when the actin cytoskeleton was disrupted by latrunculin B (Fig. 5g) or cytochalasin D (data not shown) treatment, suggesting an actin-dependent cycling of PIN3. A similar cycling pattern has been demonstrated for the homologous PIN1 protein; however, the biological meaning of this process remains unclear²¹. PIN3, in contrast to PIN1, was frequently detected in vesicles, suggesting that PIN3 cycles more rapidly or that the equilibrium of intracellular PIN3 pools is shifted more in favour of the internalized PIN3 pool.

The rapid PIN3 cycling in gravity-sensing cells (starch sheath, columella)²² raised the possibility that this cycling may provide the mechanism for rapid relocation of PIN3 after perception of an environmental stimulus. We examined the localization of the PIN3 protein in columella cells after a change in the gravity vector. Already at 2 min after root rearrangement, the originally symmetric PIN3 distribution (Fig. 5e) becomes asymmetric, with most of the PIN3 protein at the lateral side of columella cells (Fig. 5h). After 5 min the relocation of PIN3 protein appeared to be complete (Fig. 5i), and was maintained up to 20 min afterwards (data not shown). One hour after gravity stimulation the asymmetry of the PIN3 localization was observed at the tissue level. PIN3 protein was detected in an enlarged, asymmetric localization domain including columella initial and lateral root cap cells (Fig. 5j, compare with e and Fig. 4e).

Our findings provide the molecular evidence in support of the classical Cholodny–Went model for tropic growth². We show that asymmetric growth correlates with auxin efflux-dependent asymmetric auxin distribution and requires laterally localized PIN3 protein. Developmental requirements for PIN3, its localization, as well as the proposed biochemical function for PIN proteins make PIN3 a likely candidate for an efflux component of the lateral auxin transport system. The actin-dependent cycling and rapid relocation of PIN3 in cells containing gravity-sensing statoliths²², taken together with previous findings supporting a direct role of the actin cytoskeleton in gravity perception^{23,24}, imply a cellular mechanism for the regulation of asymmetric auxin distribution and differential growth. After a change of the gravity vector, actin-enmeshed statoliths²⁵ sediment, causing reorganization of the actin cytoskeleton. Thus PIN3 is redirected towards one side of the columella cells and determines direction of auxin flux, which leads to asymmetric auxin accumulation and differential growth. It is possible that tropic responses of shoots are regulated by a similar mechanism, as the loss of PIN3 or the PIN3-containing starch sheath correlate with the shoot agravitropic phenotype²⁶. □

Methods

Materials used

The PIN3 gene was identified in a bacterial artificial chromosome genomic library (<http://www.mpimp-golm.mpg.de/mpimp-map/>) with PIN1 probe (nucleotides 1–385; GenBank accession number AF089084). The full-length PIN3 cDNA was isolated from a stem cDNA library (GenBank accession number AF087818). The PIN3::GUS construct was generated by fusion of a PCR-amplified fragment (nucleotides -1,764 to -1) upstream of the ATG codon and the GUS gene. We used the following probes and primers: nucleotides 999–1449 of the PIN3 cDNA, 5'-TCCTCTCACTTCTCTCTCTCCTC-3'; 5'-TTTATTTATCTTTTCTTTGTCTCG-3'.

Phenotype analyses

Seedlings were grown as described previously¹⁰. For hypocotyl bending, 3-day-old etiolated DR5::GUS, Col-0 (Columbia ecotype) and pin3 seedlings were transferred and orientated on vertical plates with or without 10 μ M 1-N-naphthylphthalamic acid (NPA), immediately subjected to the gravity or unilateral light stimulation for 20 h stained for GUS activity, and photographed. The root gravitropism was scored in the dark in 3-day-old, light-grown seedlings, 6 h after turning the roots 135°. The quantification of responses was performed with Adobe Illustrator software. Apical hook opening was scored 50 h after germination in vertically grown, etiolated seedlings. The described defects were observed in all alleles examined (pin3-1, -2 and -3).

Expression and localization analysis

Histochemical staining for GUS activity, northern blot analysis, *in situ* hybridization, immunolocalization and immunogold electron microscopy were performed as

described^{10,12}. PIN3-specific antibodies were generated using a recombinant protein corresponding to amino acids 334–483. Affinity-purified primary anti-PIN3, fluorescein isothiocyanate-conjugated and CY3-conjugated anti-rabbit secondary antibodies (Dianova) were diluted 1:40, 1:200 and 1:600, respectively. Brefeldin A (Molecular Probes) and latrunculin B (Duchefa) treatments were performed with corresponding controls before immunolocalization as described²¹. For the gravitropism treatment, before immunolocalizations, plates with vertically grown seedlings were placed horizontally, and samples were fixed directly on plates at specific time points.

Received 17 October; accepted 10 December 2001.

1. Darwin, C. & Darwin, F. in *Darwins Gesammelte Werke Bd. 13* (Schweizerbart'sche Verlagsbuchhandlung, Stuttgart, 1881).
2. Went, F. W. Reflections and speculations. *Annu. Rev. Plant Physiol.* **25**, 1–26 (1974).
3. Hart, J. W. *Plant Tropism and Other Movements* (Unwin Hyman, London, 1990).
4. Epel, B. L., Warmbrodt, R. P. & Bandurski, R. S. Studies on the longitudinal and lateral transport of IAA in the shoots of etiolated corn seedlings. *J. Plant Physiol.* **140**, 310–318 (1992).
5. Li, Y., Hagen, G. & Guilfoyle, T. J. An auxin-responsive promoter is differentially induced by auxin gradients during tropisms. *Plant Cell* **3**, 116–1176 (1991).
6. Luschign, C., Gaxiola, R. A., Grisafi, P. & Fink, G. R. EIR1, a root specific protein involved in auxin transport, is required for gravitropism in *Arabidopsis thaliana*. *Genes Dev.* **12**, 2175–2187 (1998).
7. Friml, J. & Palme, K. Polar auxin transport—old questions and new concepts. *Plant Mol. Biol.* (in the press).
8. Palme, K. & Gälweiler, L. PIN-pointing the molecular basis of auxin transport. *Curr. Opin. Plant Biol.* **2**, 375–381 (1999).
9. Rubery, P. H. & Sheldrake, A. R. Carrier-mediated auxin transport. *Planta* **118**, 101–121 (1974).
10. Müller, A. et al. AtPIN2 defines a locus of *Arabidopsis* for root gravitropism control. *EMBO J.* **17**, 6903–6911 (1998).
11. Rashotte, A. et al. Basipetal auxin transport is required for gravitropism in roots of *Arabidopsis*. *Plant Physiol.* **122**, 481–490 (2000).
12. Gälweiler, L. et al. Regulation of polar auxin transport by AtPIN1 in *Arabidopsis* vascular tissue. *Science* **282**, 2226–2230 (1998).
13. Morris, D. A. & Thomas, A. A microautoradiographic study of auxin transport in the stem of intact pea seedlings (*Pisum sativum* L.) *J. Exp. Bot.* **29**, 147–157 (1978).
14. Björkman, T. & Cleland, R. E. The role of the epidermis and cortex in gravitropic curvature of maize roots. *Planta* **176**, 513–518 (1988).
15. Ulmasov, T., Murfett, J., Hagen, G. & Guilfoyle, T. J. Aux/IAA proteins repress expression of reporter genes containing natural and highly active synthetic auxin response elements. *Plant Cell* **9**, 1963–1971 (1997).
16. Casimiro, I. et al. Auxin transport promotes *Arabidopsis* lateral root initiation. *Plant Cell* **13**, 843–852 (2001).
17. Friml, J. et al. AtPIN4 mediates sink driven auxin gradients and patterning in *Arabidopsis* roots. *Cell* (in the press).
18. Wisman, E., Cardon, G. H., Fransch, P. & Saedler, H. The behaviour of the autonomous maize transposable element *En/Spm* in *Arabidopsis thaliana* allows efficient mutagenesis. *Plant Mol. Biol.* **37**, 989–999 (1998).
19. Ruegger, M. et al. Reduced naphthylphthalamic acid binding in the *tir3* mutant of *Arabidopsis* is associated with a reduction in polar auxin transport and diverse morphological defects. *Plant Cell* **9**, 745–757 (1997).
20. Jensen, P. J., Hangarter, R. P. & Estelle, M. Auxin transport is required for hypocotyl elongation in light-grown *Arabidopsis*. *Plant Physiol.* **16**, 455–462 (1998).
21. Geldner, N. et al. Polar auxin transport inhibitors block PIN1 cycling and vesicle trafficking. *Nature* **413**, 425–428 (2001).
22. Hasenstein, K. H. Gravisensing in plants and fungi. *Adv. Space Res.* **24**, 677–685 (1999).
23. Sievers, A., Sondag, C., Trebacz, K. & Hejnowicz, Z. Gravity induced changes in intracellular potentials in statocytes of cress roots. *Planta* **197**, 392–394 (1995).
24. Monshausen, G. B., Zieschang, H. E. & Sievers, A. Differential proton secretion in the apical elongation zone caused by gravistimulation is induced by a signal from the root cap. *Plant Cell Environ.* **19**, 1408–1414 (1996).
25. Collings, D. A., Zsuppan, G., Allen, N. S. & Blancaflor, E. B. Demonstration of prominent actin filaments in the root columella. *Planta* **212**, 392–403 (2001).
26. Fukaki, H. et al. Genetic evidence that the endodermis is essential for shoot gravitropism in *Arabidopsis thaliana*. *Plant J.* **14**, 425–430 (1998).

Acknowledgements

We thank G. Jürgens for enabling J.F. to accomplish part of this work in his laboratory; P. Tänzler and M. Sauer for technical assistance; H. Vahlenkamp for technical assistance in immunocytochemistry; M. Estelle for providing material and suggestions; T. Altman for BAC filter sets; the ADIS (Automated DNA Isolation and Sequencing) service group for DNA sequencing; ZIGIA (Center for Functional Genomics in *Arabidopsis*) for the *En* lines; and N. Geldner, T. Hamann, G. Jürgens, K. Schrick and C. Schwechheimer for comments and critical reading of the manuscript. This work was supported by a fellowship of the DAAD (J.F.), the DFG (Schwerpunktprogramm Phytohormone), the Fonds der chemischen Industrie, the European Communities Biotechnology Programs, the INCO-Copernicus Program and the European Space Agency MAP-Biotechnology Programme.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.F. (e-mail: jiri.friml@zmbp.uni-tuebingen.de) or K.P. (e-mail: palme@mpiz-koeln.mpg.de).

The non-coding *Air* RNA is required for silencing autosomal imprinted genes

Frank Sleutels*, Ronald Zwart* & Denise P. Barlow†

* Department of Molecular Genetics, The Netherlands Cancer Institute, Plesmanlaan 121, 1066CX, Amsterdam, The Netherlands

† ÖAW Institute of Molecular Biology, Billrothstrasse 11, A5020 Salzburg, Austria

In genomic imprinting, one of the two parental alleles of an autosomal gene is silenced epigenetically by a *cis*-acting mechanism^{1,2}. A bidirectional silencer for a 400-kilobase region that contains three imprinted, maternally expressed protein-coding genes (*Igf2r*/*Slc22a2*/*Slc22a3*) has been shown by targeted deletion to be located in a sequence of 3.7 kilobases^{3–5}, which also contains the promoter for the imprinted, paternally expressed non-coding *Air* RNA⁶. Expression of *Air* is correlated with repression of all three genes on the paternal allele⁵; however, *Air* RNA overlaps just one of these genes in an antisense orientation⁶. Here we show, by inserting a polyadenylation signal that truncates 96% of the RNA transcript, that *Air* RNA is required for silencing. The truncated *Air* allele maintains imprinted expression and methylation of the *Air* promoter, but shows complete loss of silencing of the *Igf2r*/*Slc22a2*/*Slc22a3* gene cluster on the paternal chromosome. Our results indicate that non-coding RNAs have an active role in genomic imprinting.

Most mammalian imprinted genes are found in clusters that also contain imprinted non-coding RNAs^{2,7}. In many cases, expression of the non-coding RNA from one parental allele correlates with repression of linked protein-coding genes on the same allele, raising the possibility that non-coding RNAs either are required for silencing or are a consequence of the silencing mechanism. Studies

of the *Igf2*/*Ins2*/*H19* imprinted gene cluster on mouse chromosome 7 have shown, however, that the *H19* non-coding RNA does not have a role in silencing the *Igf2*/*Ins2* protein-coding genes^{8,9}. Instead, imprinted expression of *Igf2*/*Ins2* occurs through a silencer element¹⁰ and an insulator element (called *H19*-DMR) close to the *H19* promoter, and imprinted expression of *H19* is thought to occur as a consequence of paternal-specific methylation of the insulator element^{11,12}. Thus, despite the prevalence of imprinted autosomal non-coding RNAs, none has been shown to be required for the silencing of flanking imprinted protein-coding genes.

We have studied the paternal-specific *Air* non-coding RNA on mouse chromosome 17 that is expressed from a promoter located in intron 2 of the *Igf2r* gene. This non-coding RNA is 108 kilobases (kb), polyadenylated but apparently unspliced, and overlaps the *Igf2r* promoter^{3,6}. Deletion of a 3.7-kb imprint control element (ICE) that contains the *Air* promoter releases paternal-specific silencing of three genes, *Slc22a2* and *Slc22a3* that are located upstream, and *Igf2r* that is located downstream^{4,5}. Thus, the 3.7-kb ICE has a bidirectional action despite the fact that *Air* overlaps only one of these genes (ref. 5; and Fig. 1a).

To test whether imprinted expression of the *Air* RNA is required for silencing, we truncated *Air* to 4% of its length by inserting a 1.2-kb polyadenylation cassette from the β -globin gene at the downstream border of the 3.7-kb ICE to yield the *Air-T* allele (Fig. 1). The modification was designed to truncate *Air* without disrupting the function of the 3.7-kb ICE, as assessed by its ability to attract a maternal-specific methylation imprint and to show paternal-specific expression of the *Air* promoter^{3,6}. Mice with a maternally inherited *Air-T* allele (*Air-T*/+, note that the maternal allele precedes the paternal one in all genotypes shown here) were identical to wild type, whereas mice with a paternally inherited *Air-T* allele (+/*Air-T*) and homozygous *Air-T* mice showed a 15% reduction in birth weight as compared with wild-type littermates (data not shown). A similar phenotype caused by biallelic expression of *Igf2r* is seen in mice with a paternally inherited deletion of the 3.7-kb ICE; these mice do not express *Air* owing to

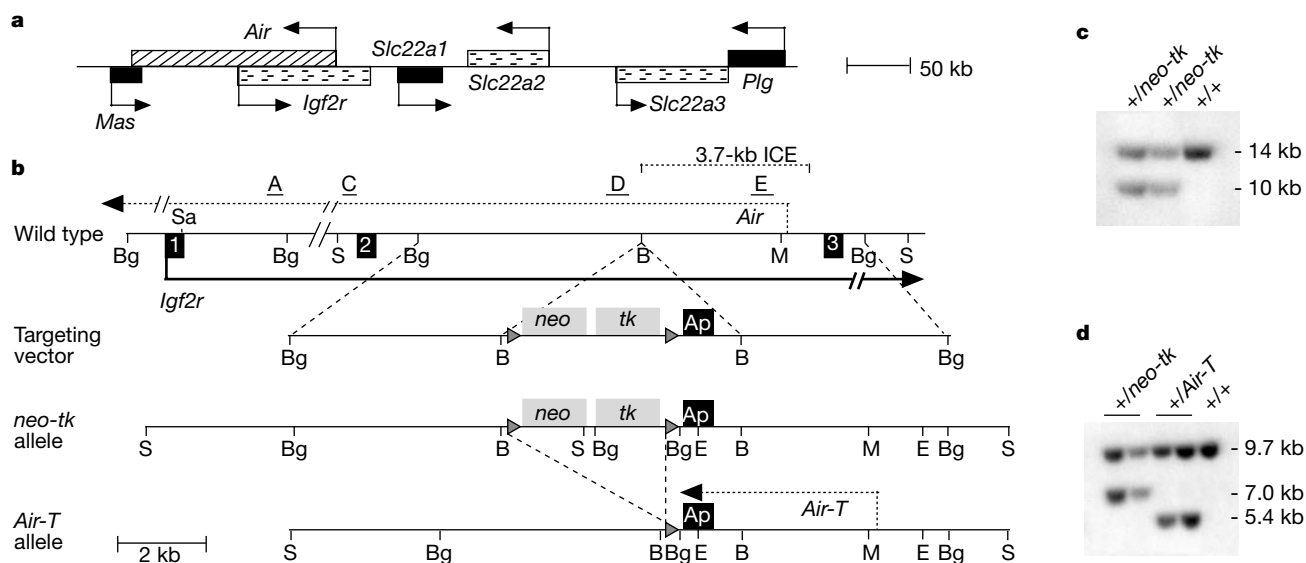


Figure 1 Generation of the *Air-T* allele. **a**, Map spanning the 400-kb imprinted cluster. Indicated are paternally expressed *Air* (hatched box), maternally expressed *Igf2r*, *Slc22a2* and *Slc22a3* (dotted boxes) and non-imprinted genes (black boxes). **b**, Wild-type allele comprises *Igf2r* (black arrow), *Igf2r* exons 1–3 (black boxes) and *Air* (dashed arrow). The locus was targeted at a *Bam*HI site bordering the 3.7-kb ICE (dotted line) that lies 3.0-kb downstream from the *Air* transcription start. The targeting vector contains a 1.2-kb rabbit β -globin gene polyadenylation cassette (Ap), and PGK–

neomycin (*neo*) and PGK–thymidine-kinase (*tk*) selection genes flanked by loxP sites (triangles). Selection genes were deleted from the *neo-tk* allele by transient Cre transfection to generate the *Air-T* allele. Probes A, C, D and E are indicated. B, *Bam*HI; Bg, *Bgl*II; E, *Eco*RI; M, *Mlu*I; Sa, *Sall*; S, *Scal*. **c**, Targeted *neo-tk* clones were identified by a 10-kb *Scal* fragment with probe C. Targeting was paternal-specific in 18 out of 18 ES clones (Supplementary Information). **d**, The *Air-T* allele was identified by a 5.4-kb *Bgl*II fragment with probe D.

deletion of the *Air* promoter⁴.

We also analysed the *Air-T* allele in reciprocal, double heterozygous crosses with mice containing a 3-cM deletion that spanned this imprinted cluster (*T^{hp}*). In a wild-type background, maternal inheritance of the *T^{hp}* allele (*T^{hp}/+*) is lethal between 15.5 and 17.5 days postcoitum (d.p.c.), owing to the absence of *Igf2r*^{4,13}. We found, however, that offspring with the *T^{hp}/Air-T* genotype were normal sized, viable and fertile, indicating that the maternal-specific *T^{hp}* phenotype had been rescued by the paternal *Air-T* allele (data not shown), as shown previously in crosses with a paternal 3.7-kb ICE deletion^{4,5}. Offspring with the reciprocal genotype (*Air-T/T^{hp}*) showed similar size, viability and fertility (data not shown) to a *+/T^{hp}* genotype.

We tested that the *Air* RNA expressed from the *Air-T* allele was truncated by using an RNase protection assay (RPA) with probes positioned across the 108 kb that is spanned by the endogenous *Air* RNA⁶ (Fig. 2a). Probes downstream from the inserted polyadenylation signal detected *Air* RNA from the paternal wild-type allele but not from the paternal *Air-T* allele, indicating the absence of stable *Air* RNA downstream from the polyadenylation cassette (Fig. 2b). We tested for polyadenylation at the inserted cassette by polymerase chain reaction with reverse transcription (RT-PCR) using a poly(dT) reverse transcription primer plus the PX6R gene-specific primer (Fig. 2a). Unexpectedly, RT-PCR generated a single 314-base-pair (bp) *Air-T*-specific fragment from a spliced, polyadenylated RNA that had used a splice donor located 53-bp downstream from the principal *Air* transcription start and a splice acceptor from the inserted exon 3 of the β -globin gene (Supplementary Information). The +53-bp splice donor has been identified in a transgene study but is not normally used in the wild-type locus; however, use

of this splice donor does not correlate with the loss of imprinting of *Igf2r/Air* in transgenes^{6,14}.

We investigated the proportion of spliced/unspliced *Air-T* RNA by RPA with probes positioned between the *Air* promoter and the inserted polyadenylation cassette (Fig. 2c). Both probe G and probe J showed that 20% unspliced *Air* RNA was present from the paternal *Air-T* allele as compared with the paternal wild-type allele (Fig. 2c). Probe K (comprising the RT-PCR fragment) detected both spliced and unspliced RNAs, and confirmed that these were present at 80% and 20%, respectively. Probe L spans the inserted polyadenylation signal and extends 404-bp downstream from the ATTTAA signal; this probe detected the same 173-bp protected fragment as found in the control LAI-313 transgene¹⁴, showing that the ATTTAA signal is used, but not larger protected fragments (Fig. 2c). Thus, Fig. 2 shows that although the *Air-T* allele generates spliced and unspliced *Air* RNAs, these are both truncated 4.2-kb downstream from the *Air* promoter and stable RNA is not detected downstream from the polyadenylation cassette.

We tested the function of the 3.7-kb ICE in the *Air-T* allele by analysing the methylation state, imprinted expression and the expression level of the *Air* promoter in *Air-T* mice as compared with wild-type littermates. Methyl-sensitive enzymes identified methylation specific to the maternal allele on the *Air-T* promoter as found on the wild-type allele (Fig. 3a). Analysis by RPA showed imprinted, paternal-specific expression of the *Air-T* RNA as found on the wild-type allele, and also showed that the *Air-T* allele generates similar levels of stable *Air* RNA as compared with the wild-type allele (Fig. 3b). Although it is possible that insertion of the polyadenylation cassette disturbed an as yet unidentified imprinting element or altered transcription/stability of the *Air-T* RNA, the

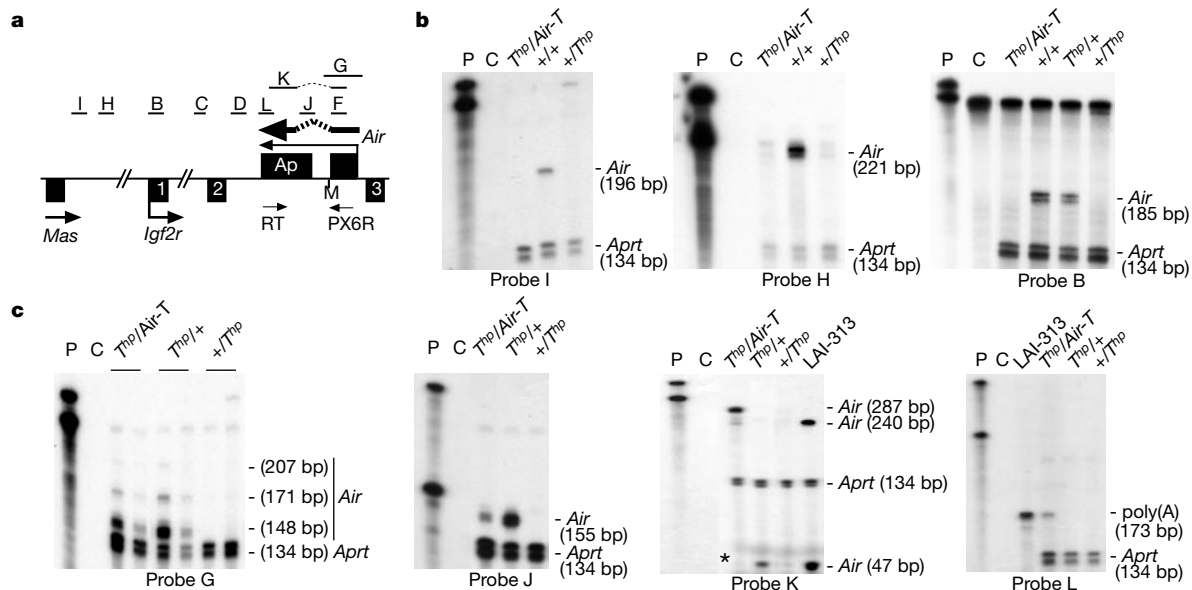


Figure 2 The *Air-T* allele truncates the *Air* RNA. **a**, Position of RNA probes (B–D, F–J, K (intron spanning), L), and RT and PX6R oligonucleotides used in RT-PCR, relative to those of *Mas*, *Igf2r* (exons 1–3) and *Air* on the *Air-T* allele. Unspliced *Air-T*, thin arrow; spliced *Air-T*, dashed arrow; *Air-T* exon 1, black box upstream from the *MluI* (M) site; polyadenylation cassette, Ap. **b**, RPA with probes B (on 11.5-d.p.c. embryo RNA), H and I (on adult heart RNA) indicates the absence of *Air* at, respectively, 28, 41 and 95 kb downstream from Ap. Identical results were obtained with probes C and D (Supplementary Information). **c**, RPA with probes G and J–L on 11.5-d.p.c. embryo RNA shows expression of truncated spliced and unspliced *Air-T* RNA. Probe G (which overlaps a +53-bp splice donor and multiple *Air* transcription start sites) protects 207, 171 and 148 bp of unspliced *Air* RNA from the *Air-T* paternal allele. However, the signal quantified by phosphorimager was reduced by 80% compared with that of the wild-type paternal allele (note the relative loading of *Air-T* and *T^{hp}/+* and that spliced fragments

protected by probe G are not visible). The signal reduction for unspliced *Air-T* is confirmed by probe J (which is located in the intron of the spliced *Air-T* RNA). Probe K is an RT-PCR fragment generated with oligonucleotides RT and PX6R (Supplementary Information) and protects 287 bp of spliced and 240 bp plus 47 bp of unspliced *Air-T* RNA (the 240-bp fragment is specific for the Ap cassette; asterisk indicates a faint 47-bp *Air* fragment). Probe K shows a ratio of 80% spliced and 20% unspliced *Air-T* RNA. The multicopy transgene LAI-313 contains the identical polyadenylation cassette¹⁴, and more clearly protects the 240- and 47-bp unspliced *Air* fragments. Probe L spans the AATAAA polyadenylation signal and protects 173 bp of *T^{hp}/Air-T* and LAI-313. The absence of larger, protected probe-L fragments confirms that both the spliced and unspliced *Air-T* RNAs are truncated at the polyadenylation signal. In **b** and **c**, parental alleles are denoted maternal/paternal, the loading control is *Apt* exon 3, input probes are indicated by P, and the tRNA control by C.

observation that the *Air-T* promoter was expressed and imprinted in a similar manner to that of the *Air* wild-type promoter indicates that function of the 3.7-kb ICE was not disturbed.

We examined the effect of the truncated *Air-T* RNA on paternal silencing of the imprinted cluster. Of the three genes in this cluster, only *Igf2r* becomes methylated when silenced on the wild-type paternal allele⁵. Methylation of *Igf2r* is completely absent from the paternal *Air-T* allele in contrast to the wild-type allele (Fig. 4a). We used RNA blots to assess imprinted expression and show (Fig. 4b, c) that all three genes, *Igf2r*, *Slc22a2* and *Slc22a3*, have a complete loss of silencing on the paternal *Air-T* allele. Maternal expression of all three genes from the *Air-T* allele was at the same level as maternal expression from the wild-type allele, indicating that the targeting event did not alter their expression.

The *Air-T* allele truncated the *Air* RNA to 4% of its length yet maintained the imprinted status and expression level of the wild-type paternal *Air* promoter (with the caveats noted above). This truncation caused a complete loss of silencing of the *Igf2r/Slc22a2/Slc22a3* cluster on the paternal chromosome. This shows that the *Air* non-coding RNA is required for *cis* repression of flanking protein-coding genes. The identification of a role for the *Air* non-coding RNA in *cis* repression contrasts with the previous observation that DNA insulator and silencer elements regulate imprinting of the *Ins2/Igf2/H19* cluster^{10–12}. Our results thus identify a new RNA-dependent mechanism in genomic imprinting.

The bidirectional action of the *Air* RNA shown by our data is unexpected because the wild-type *Air* RNA overlaps only one gene.

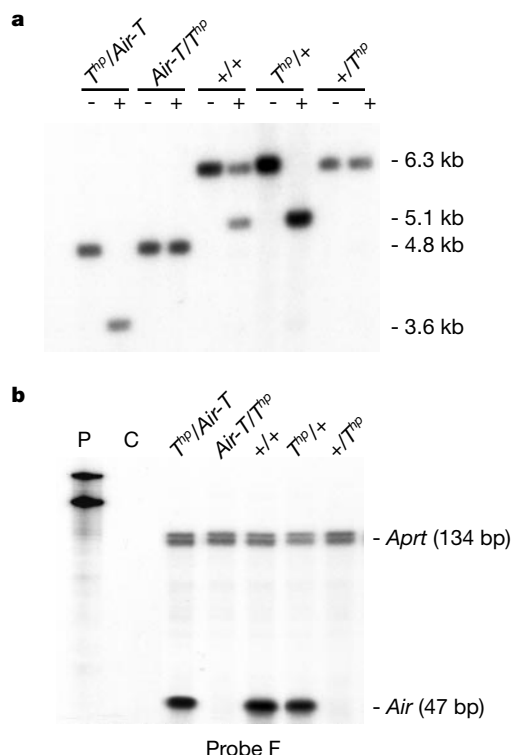


Figure 3 The *Air-T* allele is imprinted. **a**, The *Air-T* allele retains a methylation imprint on the *Air* promoter. DNA from 11.5-d.p.c. embryos cut with *EcoRI* (-) or *EcoRI/MluI* (+) was analysed with probe E. The fragments correspond to wild type methylated, 6.3 kb; wild type unmethylated, 5.1 kb; *Air-T* methylated, 4.8 kb; and *Air-T* unmethylated, 3.6 kb. Identical results were obtained with adult tail and spleen, and *SfuI* (Supplementary Information). **b**, The *Air-T* allele maintains imprinted *Air* expression. RPA was carried out on 11.5-d.p.c. embryo RNA with probe F, which detects spliced and unspliced *Air*. The paternal *Air-T* and wild-type alleles show equal *Air* expression, whereas the maternal alleles repress *Air*. Identical results were obtained with heart RNA and probes G, J and K (Supplementary Information). P, input probes; C, tRNA control.

Repression, at least of *Slc22a2* and *Slc22a3*, does not require transcriptional overlap, yet overlap with non-coding RNAs has been found at five independent imprinted loci^{3,15–18}. It is possible that the transcriptional overlap by non-coding RNAs may be involved if silencing of the non-overlapped genes were secondary to silencing of the overlapped gene. In this two-step model, the antisense non-coding RNA might repress the overlapped promoter (for example, by promoter occlusion or a form of *cis*-acting RNA interference; note that expression–competition³ is excluded by our findings) and induce a silent chromatin state that could spread bidirectionally in a limited manner and silence flanking genes.

Alternatively, if transcriptional overlap is not involved in *cis* repression, then the non-coding RNAs might function to recruit repressor proteins to the gene cluster. This latter model has parallels with X inactivation in female mammals, in which expression of the non-coding *Xist* RNA inactivates one X chromosome in *cis*¹⁹. The *Xist* RNA has features that have not been characterized as yet or that might be absent from imprinted non-coding RNAs, such as the ability to coat and spread along the length of the repressed chromosome^{20–23}. Although further testing is clearly necessary to determine whether there are mechanistic similarities at the molecular level, it is notable that X inactivation has been suggested to have evolved from a localized form of imprinting that initially

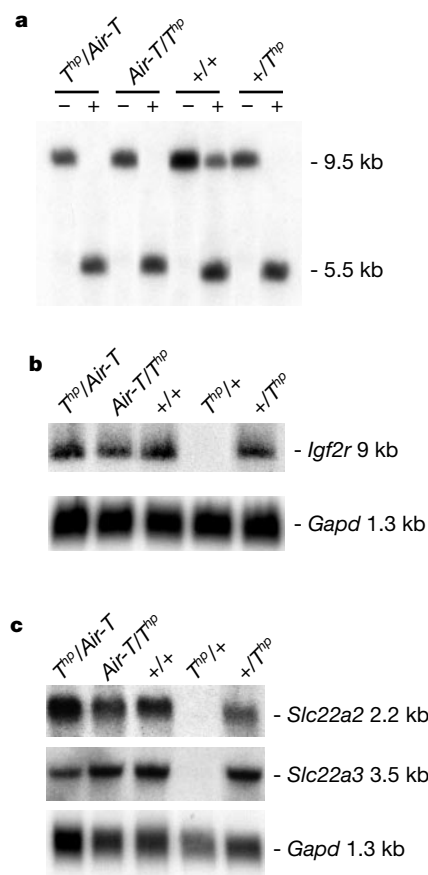


Figure 4 Loss of paternal silencing on the *Air-T* allele. **a**, The *Igf2r* promoter on the *Air-T* allele lacks paternal methylation. Tail DNA was cut with *BglII* (-) or *BglII/SalI* (+) and analysed with probe A. Methylated fragment, 9.5 kb; unmethylated fragment, 5.5 kb. Identical results were obtained with adult spleen DNA and either *NotI* or *SmaI* (Supplementary Information). **b**, Loss of paternal *Igf2r* repression on the *Air-T* allele. RNA from 11.5-d.p.c. embryos was analysed with a probe covering exons 3–6 from *Igf2r*. Identical results were obtained with adult heart RNA (Supplementary Information). **c**, Loss of paternal *Slc22a2* and *Slc22a3* repression on the *Air-T* allele. RNA from 11.5-d.p.c. placenta was analysed with cDNA probes for *Slc22a2* and *Slc22a3*. In **b** and **c**, *Gapd* is a control for RNA loading.

affected only a small region of the X chromosome that contains a dosage-sensitive gene^{24,25}. □

Methods

Generation of the *Air-T* allele

The targeting vector was constructed from a 9.6-kb *Bgl*II fragment (bp 117,878–127,575; AJ249895). A selection cassette containing *neomycin* and the thymidine kinase gene, each driven by a PGK promoter flanked by loxP sites plus a 1.2-kb fragment from the rabbit β -globin gene (bp 31,392–32,590; M18818) containing part of exon 2, complete intron 3 and complete exon 3 with the polyadenylation signal in the correct orientation for the *Air* promoter, was inserted at the *Bam*HI site in this fragment. The selection cassette was deleted by electroporation of a plasmid encoding Cre recombinase²⁶ and transient puromycin selection. We generated chimaeric mice by injecting paternally targeted *Air-T* embryonic stem (ES) cells into C57/Bl6 blastocysts²⁷. Mice carrying the *Air-T* allele were maintained on an FVB/N background.

Methylation and expression analyses

Digestion of methyl-sensitive enzymes was monitored by hybridization to mitochondrial DNA. We carried out RPA using an RPAIII kit (Ambion). Signals were quantified with a Phosphorimager (Fujix). For RT-PCR, 5 μ g heart RNA was reverse transcribed with Superscript (Gibco) using the linker-poly(dT) primer RT, 5'-CTGGGAAACAGCTATGACCATGATCGATTTTTTTTTTTTTTTT-3', and amplified with PX6R, 5'-GAAGCACAGCACCAGCAGTAC-3', for 30 cycles (94 °C, 30 s; 59 °C, 30 s; 72 °C, 120 s).

Probes for DNA analyses

We used the following probes: probe A, a 325-bp PCR fragment (bp 102,813–103,137; AJ249895); probe C, a 0.9-kb *Eco*RI/*Hin*CII fragment from EST AA592338 (ref. 6) collinear with genomic DNA; probe D, (bp 121,834–122,862; AJ249895); and probe E (bp 124,992–126,086; AJ249895).

Probes for RNA analyses

Probe B, described as probe GFP-PAIR¹⁴, is 322 bp and protects 185 bp of *Air* RNA at the *Igf2r* promoter. Probes C and D are the same as the DNA analyses probes. Probe F is 174 bp and protects 47 bp (bp 126,181–126,227; AJ249895) immediately downstream from the principal *Air* transcription start site⁶. Probe G (bp 126,086–126,293; AJ249895), described as probe MIMs1 (refs 6,14), detects unspliced (207, 171 and 148 bp) *Air* RNA fragments. This relatively large probe did not produce a clear signal for the spliced *Air-T* RNA (predicted fragments of 112, 76 and 53 bp), which should have been recognized (see probe K below). Probe H, described as probe RPA1 (ref. 6), protects 221 bp of *Air* RNA (bp 85,250–85,029; AJ249895). Probe I, described as probe MS1B8 (ref. 6) protects 196 bp. Probe J is 173 bp and protects 155 bp of *Air* RNA (bp 123,233–123,387; AJ249895). Probe K is 364 bp, contains the RT-PCR fragment made with probes RT and PX6R (Supplementary Information) and protects 287 bp of spliced and 240 and 47 bp of unspliced polyadenylated RNA. Probe L spans 558 bp (bp 32,032–32,590; M18818) of the β -globin gene (including the polyadenylation signal) and protects 173 bp of polyadenylated RNA (bp 32,032–32,204; M18818). Probe L extends 385 bp downstream from the polyadenylation signal and the absence of protected fragments longer than 173 bp indicates use of the polyadenylation signal. The *Aprt* template is a 252-bp *Xho*I/*Xba*I fragment (bp 2,165–2,417; M11310) and protects *Aprt* exon 3 (134 bp). For RNA blots, we used the following probes: for *Igf2r*, complementary DNA exons 3–6; for *Slc22a2*, bp 989–1,605 (AJ006036); for *Slc22a3* bp 1–2,766 (AF078750); for *Gapd*, complete cDNA.

Received 27 September; accepted 4 December 2001.

- Reik, W. & Walter, J. Genomic imprinting: parental influence on the genome. *Nature Rev. Genet.* **2**, 21–32 (2001).
- Sleutels, F. & Barlow, D. P. in *Homology Effects* (eds Wu, C.-t. & Dunlap, C.) (Academic, San Diego, in the press).
- Wutz, A. *et al.* Imprinted expression of the *Igf2r* gene depends on an intronic CpG island. *Nature* **389**, 745–749 (1997).
- Wutz, A. *et al.* Non-imprinted *Igf2r* expression decreases growth and rescues the Tme mutation in mice. *Development* **128**, 1881–1887 (2001).
- Zwart, R., Sleutels, F., Wutz, A., Schinkel, A. H. & Barlow, D. P. Bidirectional action of the *Igf2r* imprint control element on upstream and downstream imprinted genes. *Genes Dev.* **15**, 2361–2366 (2001).
- Lyle, R. *et al.* The imprinted antisense RNA at the *Igf2r* locus overlaps but does not imprint *Mas1*. *Nature Genet.* **25**, 19–21 (2000).
- Beechey, C. V., Cattanach, B. M. & Selley, R. L. Mouse Imprinting Data and References. *MRC Mammalian Genetics Unit* [online] (<http://www.mgu.har.mrc.ac.uk/imprinting/imprinting.html>) (2000).
- Schmidt, J. V., Levers, J. M. & Tilghman, S. M. Enhancer competition between H19 and *Igf2* does not mediate their imprinting. *Proc. Natl Acad. Sci. USA* **96**, 9733–9738 (1999).
- Hark, A. T. *et al.* CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/*Igf2* locus. *Nature* **405**, 486–489 (2000).
- Constancia, M. *et al.* Deletion of a silencer element in *Igf2* results in loss of imprinting independent of H19. *Nature Genet.* **26**, 203–206 (2000).
- Bell, A. C. & Felsenfeld, G. Methylation of a CTCF-dependent boundary controls imprinted expression of the *Igf2* gene. *Nature* **405**, 482–485 (2000).
- Reik, W. & Murrell, A. Genomic imprinting. Silence across the border. *Nature* **405**, 408–409 (2000).
- Wang, Z. Q., Fun, M. R., Barlow, D. P. & Wagner, E. F. Regulation of embryonic growth and lysosomal targeting by the imprinted *Igf2/Mpr* gene. *Nature* **372**, 464–467 (1994).
- Sleutels, F. & Barlow, D. P. Investigation of elements sufficient to imprint the mouse *Air* promoter. *Mol. Cell. Biol.* **21**, 5008–5017. (2001).

- Wroe, S. F. *et al.* An imprinted transcript, antisense to *Nesp*, adds complexity to the cluster of imprinted genes at the mouse *Gnas* locus. *Proc. Natl Acad. Sci. USA* **97**, 3342–3346 (2000).
- Lee, Y. J. *et al.* *Mit1/Lb9* and *Copg2*, new members of mouse imprinted genes closely linked to *Peg1/Mest1*. *FEBS Lett.* **472**, 230–234 (2000).
- Rougeulle, C., Cardoso, C., Fontes, M., Colleaux, L. & Lalande, M. An imprinted antisense RNA overlaps *UBE3A* and a second maternally expressed transcript. *Nature Genet.* **19**, 15–16 (1998).
- Smilnich, N. J. *et al.* A maternally methylated CpG island in *KvLQT1* is associated with an antisense paternal transcript and loss of imprinting in Beckwith-Wiedemann syndrome. *Proc. Natl Acad. Sci. USA* **96**, 8064–8069 (1999).
- Avner, P. & Heard, E. X-chromosome inactivation: counting, choice and initiation. *Nature Rev. Genet.* **2**, 59–67 (2001).
- Clemson, C. M., McNeil, J. A., Willard, H. F. & Lawrence, J. B. XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure. *J. Cell Biol.* **132**, 259–275 (1996).
- Lee, J. T., Strauss, W. M., Dausman, J. A. & Jaenisch, R. A 450 kb transgene displays properties of the mammalian X-inactivation center. *Cell* **86**, 83–94 (1996).
- Sheardown, S. A. *et al.* Stabilization of Xist RNA mediates initiation of X chromosome inactivation. *Cell* **91**, 99–107 (1997).
- Wutz, A. & Jaenisch, R. A shift from reversible to irreversible X inactivation is triggered during ES cell differentiation. *Mol. Cell* **5**, 695–705 (2000).
- Lyon, M. F. Imprinting and X-chromosome inactivation. *Results Probl. Cell Differ.* **25**, 73–90 (1999).
- Graves, J. A. Mammals that break the rules: genetics of marsupials and monotremes. *Annu. Rev. Genet.* **30**, 233–260 (1996).
- O'Gorman, S., Dagenais, N. A., Qian, M. & Marchuk, Y. Protamine-Cre recombinase transgenes efficiently recombine target sequences in the male germ line of mice, but not in embryonic stem cells. *Proc. Natl Acad. Sci. USA* **94**, 14602–14607 (1997).
- Hogan, B. L. M., Beddington, R. S. P., Costantini, F. & Lacy, E. *Manipulating the Mouse Embryo* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1994).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com>).

Acknowledgements

We thank K. van Veen, K. van het Wout, P. Krimpenfort for help in generating mice; S. Greven, T. Maidment and N. Bosnie for care of the mice; A. Berns, H. te Riele, M. van Lohuizen, R. Beijersbergen, P. Borst and A. Frischau for comments; and A. Berns for help and encouragement. This research was supported by the Dutch Cancer Society (KWF).

Correspondence and requests for materials should be addressed to D.P.B. (e-mail: dbarlow@imb.oeaw.ac.at).

Structural basis for antagonist-mediated recruitment of nuclear co-repressors by PPAR α

H. Eric Xu, Thomas B. Stanley, Valerie G. Montana, Millard H. Lambert, Barry G. Shearer, Jeffery E. Cobb, David D. McKee, Cristin M. Galardi, Kelli D. Plunket, Robert T. Nolte, Derek J. Parks, John T. Moore, Steven A. Kliewer, Timothy M. Willson & Julie B. Stimmel

Nuclear Receptor Discovery Research, GlaxoSmithKline, Research Triangle Park, North Carolina 27709, USA

Repression of gene transcription by nuclear receptors is mediated by interactions with co-repressor proteins such as SMRT and N-CoR^{1,2}, which in turn recruit histone deacetylases to the chromatin^{3–5}. Aberrant interactions between nuclear receptors and co-repressors contribute towards acute promyelocytic leukaemia and thyroid hormone resistance syndrome^{6–8}. The binding of co-repressors to nuclear receptors occurs in the unliganded state, and can be stabilized by antagonists⁹. Here we report the crystal structure of a ternary complex containing the peroxisome proliferator-activated receptor- α ligand-binding domain bound to the antagonist GW6471 and a SMRT co-repressor motif. In this structure, the co-repressor motif adopts a three-turn α -helix that prevents the carboxy-terminal activation helix (AF-2) of the receptor from assuming the active conformation. Binding of

the co-repressor motif is further reinforced by the antagonist, which blocks the AF-2 helix from adopting the active position. Biochemical analyses and structure-based mutagenesis indicate that this mode of co-repressor binding is highly conserved across nuclear receptors.

The peroxisome proliferator-activated receptor- α (PPAR α) regulates hepatic fatty acid metabolism and mediates the effects of fibrate lipid-lowering drugs¹⁰. We recently described the L-tyrosine analogue GW409544 as a potent PPAR α agonist¹¹. The crystal structure of GW409544 bound to PPAR α reveals that the carboxylic acid group of GW409544 makes a direct hydrogen bond with Y464 on the C-terminal AF-2 helix, and this interaction stabilizes the receptor in the active conformation¹¹. On the basis of these observations, we designed a PPAR α antagonist by substituting an ethyl amide group for the acid group of GW409544 to disrupt the hydrogen bond with Y464 (Fig. 1a). The resulting compound, GW6471, is a potent PPAR α antagonist. In a cell-based reporter assay, GW6471 completely inhibited GW409544-induced activation of PPAR α with a median inhibitory concentration (IC₅₀) of 0.24 μ M (Fig. 1b).

The transcriptional activity of nuclear receptors is a function of their interactions with coactivators such as SRC-1 (steroid receptor coactivator-1)¹² or CBP (CREB-binding protein), or co-repressors

such as N-CoR (nuclear co-repressor) and SMRT (silencing mediator for retinoid and thyroid hormone receptors). Coactivators contain an LXXLL motif that forms two turns of α -helix and binds into a hydrophobic cleft on the surface of the receptor. In the PPAR α /GW409544/SRC-1 structure, the SRC-1 coactivator motif is secured by a charge clamp between residues on the AF-2 helix and helix 3, similar to coactivator complexes with oestrogen receptor¹³, thyroid receptor¹⁴, retinoid X receptor¹⁵ and PPAR γ ^{15,16}. As nuclear receptor antagonists decrease binding of coactivators and increase binding of co-repressors⁹, we examined the effect of GW6471 on the interaction of PPAR α with peptides derived from coactivator and co-repressor motifs^{17–21}. GW6471 disrupted the interactions between PPAR α and coactivator motifs derived from SRC-1 or CBP, but promoted the binding of the co-repressor motifs from SMRT or N-CoR²² (Fig. 1c). The potency of GW6471 for displacement of coactivators and recruitment of co-repressors matched the IC₅₀ value for PPAR α antagonism in cell-based assays. The binding affinity of the SMRT or the N-CoR motifs to the PPAR α ligand-binding domain (LBD) was enhanced fivefold by GW6471 (Fig. 1d). GW6471 also enhanced the interactions of PPAR α with the N-CoR or the SMRT motifs in a mammalian two-hybrid assay (Fig. 1e). These results demonstrate that GW6471 induces a PPAR α LBD conformation that interacts efficiently with co-repressors.

Both SMRT and N-CoR contain multiple receptor-interacting motifs^{19–21}. To understand the molecular basis of co-repressor recruitment, we determined the crystal structure of a ternary complex containing the PPAR α LBD bound to GW6471 and a peptide derived from the C-terminal receptor-interacting motif of SMRT (Figs 2a and 3a). The refined structure contains four PPAR α /GW6471/SMRT complexes (see Methods and Table 1), and reveals a clear electron density for the core co-repressor motif and the antagonist GW6471 (Fig. 2b, c). The PPAR α LBD contains 13 α -helices and 4 small β -strands that fold into a typical nuclear receptor helical sandwich. The SMRT motif adopts a three-turn α -helix and docks into a hydrophobic groove formed by helices 3, 3', 4 and 5 (Fig. 2a). Superposition of the agonist- and antagonist-bound PPAR α (Fig. 2d) reveals that the co-repressor-binding site partly overlaps with the coactivator-binding site. The additional helical turn in the co-repressor motif extends into the space that is left by the repositioning of the AF-2 helix, and prevents this helix from folding back into its active conformation.

GW6471 adopts a U-shaped configuration and wraps around C276 of helix 3 (Fig. 2c). The modified amide head group of GW6471 adopts a configuration that is no longer able to form a hydrogen bond with Y464 on the AF-2 helix. Instead, this larger head group extends into the space that is normally occupied by the side chain of Y464, and pushes the AF-2 helix out of the agonist-bound position (Fig. 2e). Unlike the previously determined antago-

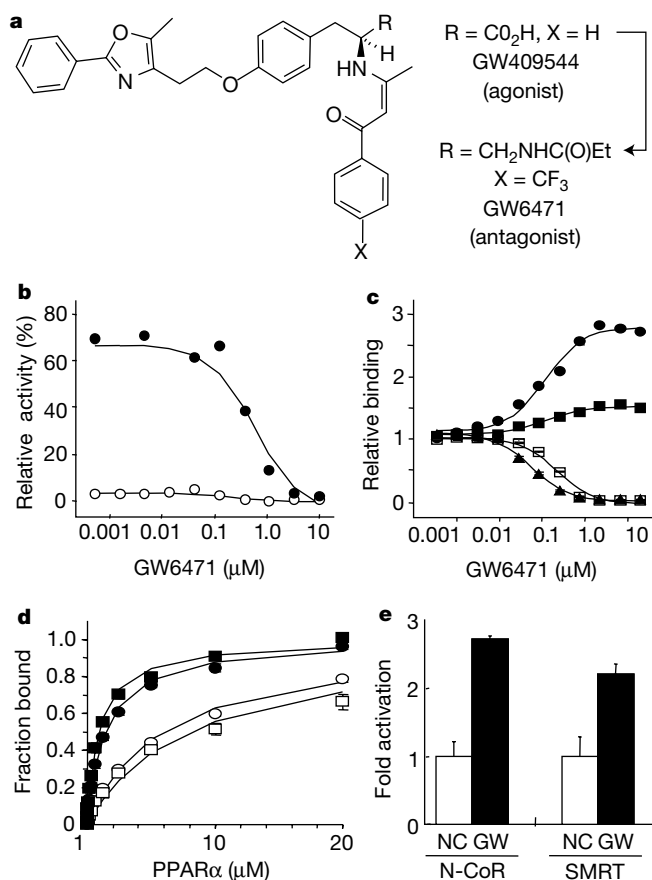


Figure 1 GW6471 recruits co-repressors to PPAR α . **a**, Chemical structures of GW409544 and GW6471. **b**, Effect of GW6471 on activation of PPAR α in the presence (filled circles) or absence (open circles) of 10 nM GW409544. **c**, Fluorescence energy transfer assays for the effect of GW6471 on peptide binding to PPAR α . SMRT, filled circles; N-CoR, filled squares; CBP, filled triangles; SRC-1, open squares. **d**, Binding of SMRT (squares) or N-CoR (circles) motifs to PPAR α in the presence (filled) or absence (open) of GW6471, as measured by fluorescence polarization. **e**, Effect of GW6471 on binding of SMRT or N-CoR motifs to PPAR α in a mammalian two-hybrid assay. NC, no compound; GW, 10 μ M GW6471. The results are the average of three experiments, with error bars showing s.d.

Table 1 Statistics of crystallographic data and structures

Crystals	PPAR α /SMRT complex	PPAR α /SRC-1 complex
Space group	<i>P</i> 2 ₁ 2 ₁ 2 ₁	<i>P</i> 2 ₁ 2 ₁ 2
Resolution (Å)	30.0–3.0	20.0–1.8
Unique reflections	29,380	30,147
Completeness (%)	99.0	99.4
<i>I</i> / σ	23.3 (2.0)	39.5 (5.2)
<i>R</i> _{sym} (%)	6.8	4.5
Refinement statistics		
<i>R</i> _{factor} (%)	25.7	21.9
<i>R</i> _{free} (%)	29.0	26.7
r.m.s.d. bond lengths (Å)	0.008	0.007
r.m.s.d. bond angles (deg)	1.956	1.561
Total non-hydrogen atoms	9,615	2,659

r.m.s.d. is the root-mean-square deviation from ideal geometry.

*R*_{sym} = $\sum |F_o - I| / \sum I$.

*R*_{factor} = $\sum |F_o - F_c| / \sum F_o$, where *F*_o and *F*_c are observed and calculated structure factors. *R*_{free} was calculated from a randomly chosen 10% of reflections excluded from refinement, and *R*_{factor} was calculated for the remaining 90% of reflections.

nist-bound oestrogen-receptor structures¹³, the AF-2 helix does not occupy the coactivator-binding groove, but is loosely packed against helix 3 (Fig. 2a, d). The PPAR α /SMRT complex reveals that removal of the AF-2 helix from its active position destroys the integrity of the charge clamp but leaves sufficient space to accommodate the additional helical turn of the co-repressor motif.

Although there is overlap in the co-repressor- and coactivator-binding modes, there are numerous differences in their binding interfaces. First, the three helical turns in the co-repressor motif result in a larger interaction interface with PPAR α than the two helical turns of the coactivator motif. The SMRT LXXXIXXXL motif covers 736 Å² of Connolly surface of PPAR α , whereas the SRC-1 LXXLL motif buries only 478 Å². The larger co-repressor interaction surface may account for the preferential binding of co-repressor over coactivator to PPAR α in the presence of GW6471. Second, the co-repressor helix is anchored to PPAR α by three hydrogen bonds between the C-terminal carbonyls and the conserved K292 from helix 3, which also caps the coactivator helix¹¹. Compared with the coactivator helix, the co-repressor motif pivots ~15 degrees about its C-terminal end into the void left behind by displacement of the AF-2 helix, bringing the central isoleucine residue (I + 5 in Fig. 3a) 2–3 Å closer to helices 3, 4 and 5 (Fig. 2d, e). The unexpected binding mode of the co-repressor motif provides a structural explanation for the sequence differences between co-repressor and coactivator motifs. The first and the second leucines of the coactivator LXXLL motif are equivalent in the positions of the central isoleucine (+5) and the A + 8 of the co-repressor motif in the superposition of their respective structures. However, the larger leucine side chains at positions +5 and +8

prevent the coactivator LXXLL motif binding tightly to the receptor in the absence of the AF-2 helix, like the co-repressor motif. This is in agreement with previous studies, which found that replacing residue +5 and +8 with leucines in the co-repressor motif results in a non-functional co-repressor²¹. The third difference between co-repressor and coactivator binding is that the three-turn helix of the co-repressor motif deviates from a regular α -helix, which allows the residues of L + 1, I + 5 and L + 9 to align on the same face of the helix (Figs 2f and 3b), and to make the core hydrophobic interactions with the receptor. Besides these core interactions, the co-repressor motif makes additional interactions with the receptor through peripheral residues composed by I + 4, A + 8, E + 2 and R + 6, which are oriented at angles of ~90° to either side of the central interface (Fig. 3b). On one side, the side chains of I + 4 and A + 8 are partially exposed and make additional hydrophobic contacts along one ridge of the binding groove. On the opposite side, residues E + 2 and R + 6 form a strong intramolecular H-bond, which positions these residues so as to make a pair of intermolecular hydrogen bonds with K310 and N303 of helix 4. This hydrogen bond network is further strengthened by additional hydrophobic interactions between these two side chains and L302 and V306, which form the ridge on this side of the groove.

We next examined the receptor–co-repressor interface by alanine-scanning mutations of the co-repressor motif (Fig. 3c). Replacement of any of the three core interface residues (L + 1, I + 5 and L + 9) with alanine markedly reduced co-repressor binding. Mutating the peripheral interface residues (E + 2, I + 4, R + 6 and M + 10) also decreased co-repressor binding. The interaction profile of these mutated co-repressor motifs was very

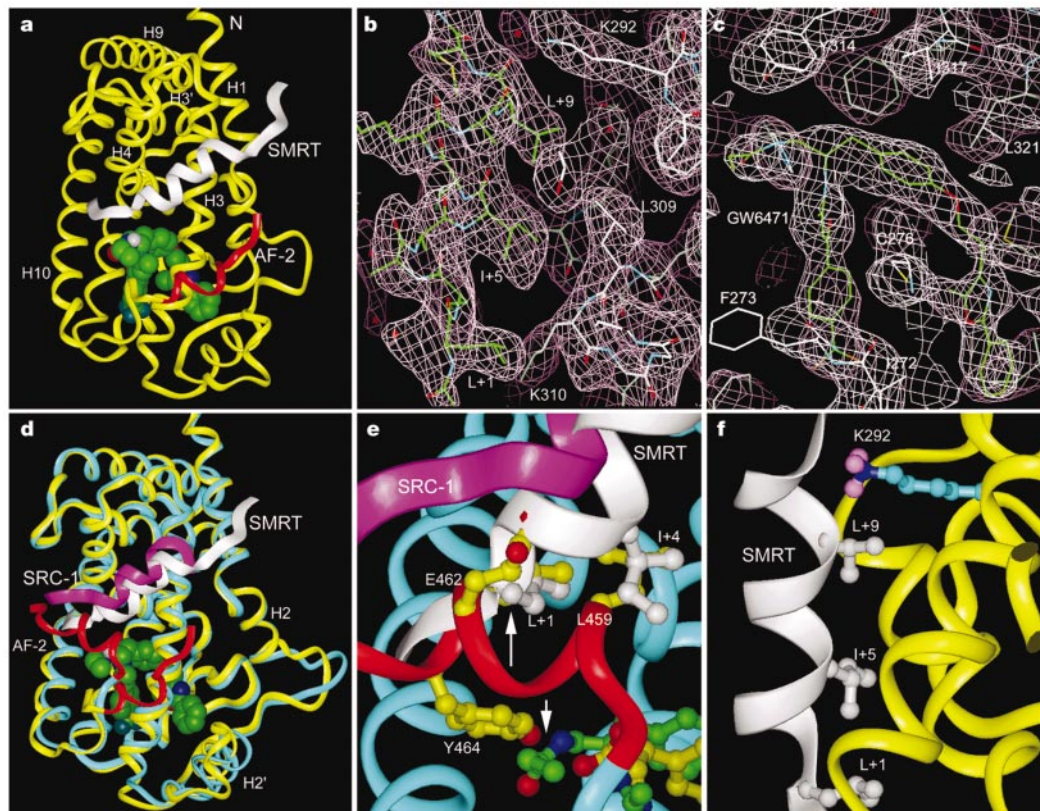


Figure 2 Structure of the PPAR α /GW6471/SMRT complex. An overview of the ternary complex. **a**, PPAR α LBD, yellow; AF-2 helix, red; SMRT motif, white; GW6471, green spheres. **b**, **c**, Electron density maps of the PPAR α –SMRT interface and the antagonist GW6471. The maps are calculated with F_0 coefficient and contoured at 1 σ . The carbon atoms of SMRT and GW6471 are in green. **d**, Overlay of the PPAR α /GW6471/SMRT and

the PPAR α /GW409544/SRC-1 complexes. PPAR α LBD, yellow in SMRT complex or blue in SRC-1 complex; SRC-1, purple; SMRT, white. **e**, Superposition of SMRT peptide and GW6471 with the PPAR α /GW409544/SRC-1 complex. The short arrow indicates interference between GW6471 and Y464; the long arrow indicates interference between SMRT and the active AF-2 helix. **f**, Interface between PPAR α (yellow) and SMRT (white).

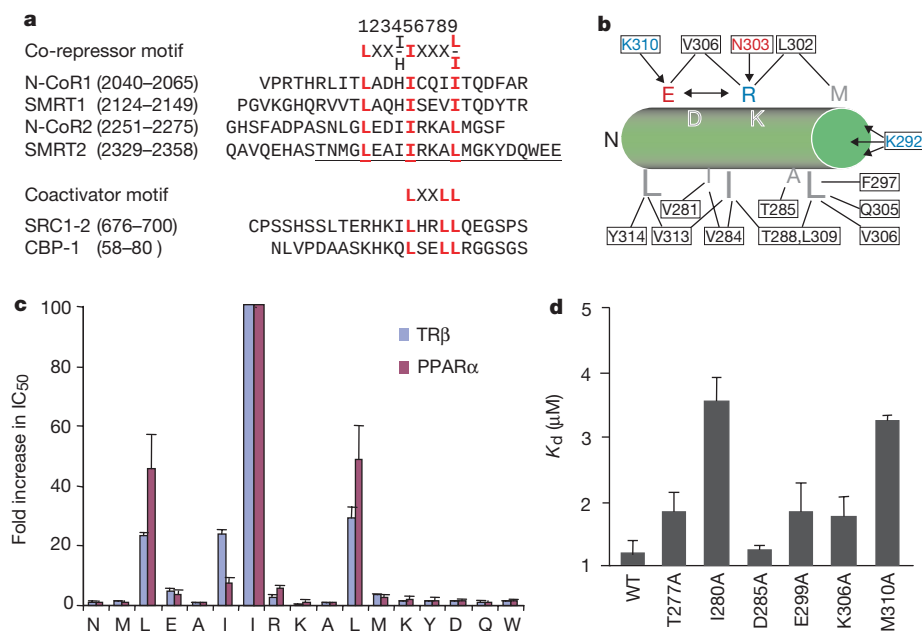


Figure 3 Conservation of co-repressor binding to nuclear receptors. **a**, Co-repressor and coactivator motifs. The underlined sequence was used in crystallography. **b**, Interactions between the SMRT motif (green cylinder) and PPAR α (boxed residues). Residue colours: H-bond donors, blue; acceptors, red; hydrophobic contacts, grey (SMRT) and black (PPAR α). Arrows indicate hydrogen bonds; lines indicate van der Waals contacts.

similar to the one for TR β (thyroid hormone receptor- β) (Fig. 3c), demonstrating a similar mode of co-repressor binding to these receptors. This idea is further supported by structure-based sequence alignment, which reveals that the residues involved in co-repressor binding are highly conserved across the nuclear receptor family (data not shown). Nine out of the fourteen residues that contact the co-repressor domain in PPAR α were previously mutated in TR β ^{19–21}. Each of these mutations decreased the interaction between TR β and the co-repressor motif. We mutated the remaining five contact residues in TR β and showed that these changes also result in decreases in co-repressor interaction (Fig. 3d). These results suggest that the co-repressor interaction interface observed in PPAR α is conserved in other nuclear receptors.

The structure of the PPAR α /GW6471/SMRT ternary complex demonstrates that the co-repressor motif adopts an irregular three-turn helix that fits tightly into a receptor groove, which also serves as the coactivator-binding site. Nuclear receptors distinguish co-repressors from coactivators by the length of their helical interaction motifs, which depends on the AF-2 conformation. In the presence of an antagonist, the AF-2 helix is blocked from assuming the active conformation, resulting in a larger pocket that can accommodate the three-turn α -helix of the co-repressor LXXXIXXXL motif. The high degree of conservation of the co-repressor and coactivator interaction interfaces suggests that this is a model that applies to many of the nuclear receptors. □

Methods

Protein preparation

The human PPAR α LBD (amino acids 198–468; GenBank accession number S74349) tagged with MKKHHHHHHHLPGRG (thrombin cleavage site underlined) was expressed from the T7 promoter of the pRSETA vector (Invitrogen). The wild-type and mutated LBDs of TR β were also expressed from pRSETA, but were tagged with MKKGHHHHHHHG. The proteins were purified as described²³, and the tag was removed from the PPAR α LBD by incubation with thrombin (200:1 molar ratio) overnight at 4 °C. The PPAR α /GW6471/SMRT complex was prepared by addition of a twofold molar excess of the SMRT2 peptide (sequence referred to in Fig. 3a) and fivefold molar excess of the antagonist, concentrated to ~9.0 mg ml⁻¹, and then aliquoted and stored at ~80 °C.

c, Alanine scan of the SMRT motif. Binding of peptides to PPAR α and TR β was measured by fluorescence polarization. IC₅₀ values for the wild-type SMRT motif are 8.0 ± 3.4 μ M and 6.7 ± 2.3 μ M for PPAR α and TR β , respectively. **d**, Affinities of mutated TR β LBDs for the SMRT2 motif as measured by fluorescence polarization. WT, wild type. Error bars are the s.d. of three experiments.

Crystalization and data collection

The PPAR α /GW6471/SMRT crystals were grown at room temperature in hanging drops containing 1 μ l of the above protein-ligand solutions and 1 μ l of well buffer (5.5–9.6% polyethylene glycol (PEG) 35K, 50 mM di-ammonium hydrogen citrate, pH 4.9, 50 mM 1,3 bis-tris-propane (BTP), pH 7.0, 10% MPD). Crystals grew to a size of 100–200 μ m within several weeks. Data collections were performed with crystals that were transiently mixed with the well buffer containing an additional 10% 2-methyl-2,4-pentanediol (MPD), and then flash frozen in liquid nitrogen.

The crystals formed in the $P2_12_12_1$ space group, with $a = 103.4$ Å, $b = 112.8$ Å, $c = 123.9$ Å. Each asymmetric unit contained four PPAR α LBDs, four SMRT co-repressor peptides and four GW6471 antagonist molecules with a solvent content of 50%. Crystals were screened with a Rigaku R-Axis II detector and data were collected with a MAR CCD (charge-coupled device) detector at 17-ID in the facilities of the Industrial Macromolecular Crystallography Association (IMCA), at the Advanced Photon Source. We processed the observed reflections with the Denzo and Scalepack programs in the HKL2000 package²⁴.

Structure determination and refinement

The PPAR α /GW6471/SMRT structure was determined by molecular replacement methods with the AmoRe program²⁵ using a 1.8 Å structure of a PPAR α /SRC-1 complex as the initial model (Table 1). The best-fit solution generated with AmoRe gave a correlation coefficient of 36.1% and an R -factor of 46.9%. The phases from the molecular replacement were extensively refined with solvent flattening, histogram matching, and fourfold non-crystallographic symmetry (NCS) averaging as implemented in the DM program²⁶. The electron density maps were further improved with multiple crystal averaging of three PPAR α /SMRT data sets by the DMMULTI program. We applied a negative b -factor to the data to enhance the feature of the electron density for the protein side chains²⁷. Multiple cycles of manual model building were performed with QUANTA (Molecular Simulations). Structure refinements were proceeded with CNX²⁸, using the maximum likelihood target and NCS constraints, which was relaxed in the final stages of refinement. The statistics of the structure and data sets are summarized in Table 1.

Functional assays

Cell-based assays for GW6471 as an antagonist were performed in CV-1 cells using the (UAS)₅-tk-SAP reporter and the human PPAR α -GAL4 chimaeras as described previously¹¹. We carried out transfections using lipofectamine (Life Technologies) with a β -galactosidase vector as a normalization control. The mammalian two-hybrid assays were performed according to the published procedures²⁹. Transfection mixes included 8 ng (UAS)₅-tk-CAT reporter plasmid, 8 ng VP16-human PPAR α expression plasmid, 2 ng of either GAL4-N-CoR or GAL4-SMRT expression plasmid, 25 ng β -galactosidase expression plasmid as internal control, and 35 ng carrier plasmid. The synthesis of GW6471 and GW409544 will be described elsewhere.

Binding assays

The effects of GW6471 on the interaction of coactivator and co-repressor peptides with

PPAR α were determined by chemical-mediated fluorescence energy transfer assays using the AlphaScreen Technology from Packard BioScience³⁰. The experiments were conducted with 5 nM PPAR α LBD of biotinylated peptide containing individual motifs (Fig. 3a), following the manufacturer's instructions for the hexahistidine detection kit in a buffer containing 50 mM MOPS, pH 7.4, 50 mM NaF, 0.05 mM CHAPS, 0.1 mg ml⁻¹ bovine serum albumin, and 10 mM dithiothreitol (DTT). The binding signals were detected with the increasing concentrations of GW6471, and the results from four repeated experiments were normalized as a percentage of the binding in the absence of GW6471.

The effects of GW6471 on the affinity of the SMRT or N-CoR peptides with purified PPAR α LBD were determined by fluorescence polarization in a buffer containing 10 mM HEPES, pH 7.4, 0.15 M NaCl, 3 mM EDTA, 0.005% polysorbate-20, 5 mM DTT and 2.5% DMSO. Varied concentration of PPAR α LBD in the presence or absence of 40 μ M GW6471 were incubated at room temperature with 10 nM of a fluorescein-labelled peptide of N-CoR2 or SMRT2 (Fig. 3a). The fluorescence polarization values for each concentration of receptor were determined using a BMG PolarStar Galaxy fluorescence reader with 485 nm excitation and 520 nm emission filters. The apparent dissociation constant (K_d) values were determined by the binding curves derived from a nonlinear least-squares-fit of the data for a simple 1:1 interaction.

Mutational analysis of the SMRT co-repressor motif interaction with the PPAR α and TR β LBDs was also performed by fluorescence polarization. To determine the importance of each amino acid in the SMRT motif for binding to nuclear receptors, SMRT peptides with alanine substitution at each position were added to inhibit the binding of 1 μ M TR β LBD or 2 μ M PPAR α to the fluorescent N-CoR2 peptide. For the PPAR α experiments we added 10 μ M GW6471. The inhibition curves were constructed and IC₅₀ values were determined by nonlinear least-squares-fit of the data to a simple 1:1 interaction.

Received 8 November; accepted 12 December 2001.

- Horlein, A. J. *et al.* Ligand-independent repression by the thyroid hormone receptor mediated by a nuclear receptor co-repressor. *Nature* **377**, 397–404 (1995).
- Chen, J. D. & Evans, R. M. A transcriptional co-repressor that interacts with nuclear hormone receptors. *Nature* **377**, 454–457 (1995).
- Nagy, L. *et al.* Nuclear receptor repression mediated by a complex containing SMRT, mSin3A, and histone deacetylase. *Cell* **89**, 373–380 (1997).
- Hassig, C. A., Fleischer, T. C., Billin, A. N., Schreiber, S. L. & Ayer, D. E. Histone deacetylase activity is required for full transcriptional repression by mSin3A. *Cell* **89**, 341–347 (1997).
- Laherty, C. D. *et al.* Histone deacetylases associated with the mSin3 corepressor mediate mad transcriptional repression. *Cell* **89**, 349–356 (1997).
- Hong, S. H., David, G., Wong, C. W., Dejean, A. & Privalsky, M. L. SMRT corepressor interacts with PLZF and with the PML-retinoic acid receptor alpha (RAR α) and PLZF-RAR α oncoproteins associated with acute promyelocytic leukemia. *Proc. Natl Acad. Sci. USA* **94**, 9028–9033 (1997).
- Grignani, F. *et al.* Fusion proteins of the retinoic acid receptor- α recruit histone deacetylase in promyelocytic leukaemia. *Nature* **391**, 815–881 (1998).
- Yoh, S. M., Chatterjee, V. K. & Privalsky, M. L. Thyroid hormone resistance syndrome manifests as an aberrant interaction between mutant T3 receptors and transcriptional corepressors. *Mol. Endocrinol.* **11**, 470–480 (1997).
- Jackson, T. A. *et al.* The partial agonist activity of antagonist-occupied steroid receptors is controlled by a novel hinge domain-binding coactivator L7/SPA and the corepressors N-CoR or SMRT. *Mol. Endocrinol.* **11**, 693–705 (1997).
- Issemann, I. & Green, S. Activation of a member of the steroid hormone receptor superfamily by peroxisome proliferators. *Nature* **347**, 645–650 (1990).
- Xu, H. E. *et al.* Structural determinants of ligand binding selectivity between the peroxisome proliferator-activated receptors. *Proc. Natl Acad. Sci. USA* **98**, 13919–13924 (2001).
- Onate, S. A., Tsai, S. Y., Tsai, M. J. & O'Malley, B. W. Sequence and characterization of a coactivator for the steroid hormone receptor superfamily. *Science* **270**, 1354–1357 (1995).
- Shiau, A. K. *et al.* The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell* **95**, 927–937 (1998).
- Darimont, B. D. *et al.* Structure and specificity of nuclear receptor–coactivator interactions. *Genes Dev.* **12**, 3343–3356 (1998).
- Gampe, R. T. Jr *et al.* Asymmetry in the PPAR γ /RXR α crystal structure reveals the molecular basis of heterodimerization among nuclear receptors. *Mol. Cell* **5**, 545–555 (2000).
- Nolte, R. T. *et al.* Ligand binding and co-activator assembly of the peroxisome proliferator-activated receptor- γ . *Nature* **395**, 137–143 (1998).
- Heery, D. M., Kalkhoven, E., Hoare, S. & Parker, M. G. A signature motif in transcriptional co-activators mediates binding to nuclear receptors. *Nature* **387**, 733–736 (1997).
- Yang, W., Rachez, C. & Freedman, L. P. Discrete roles for peroxisome proliferator-activated receptor gamma and retinoid X receptor in recruiting nuclear receptor coactivators. *Mol. Cell Biol.* **20**, 8008–8017 (2000).
- Nagy, L. *et al.* Mechanism of corepressor binding and release from nuclear hormone receptors. *Genes Dev.* **13**, 3209–3216 (1999).
- Hu, X. & Lazar, M. A. The CoRNR motif controls the recruitment of corepressors by nuclear hormone receptors. *Nature* **402**, 93–96 (1999).
- Perissi, V. *et al.* Molecular determinants of nuclear receptor–corepressor interaction. *Genes Dev.* **13**, 3198–3208 (1999).
- Zhou, G. *et al.* Nuclear receptors have distinct affinities for coactivators: characterization by fluorescence resonance energy transfer. *Mol. Endocrinol.* **12**, 1594–1604 (1998).
- Xu, H. E. *et al.* Molecular recognition of fatty acids by peroxisome proliferator-activated receptors. *Mol. Cell* **3**, 397–403 (1999).
- Otwinowski, Z. & Minor, W. in *Macromolecular Crystallography* (eds Carter, J. C. W. & Sweet, R. M.) 307–326 (Academic, New York, 1997).
- Navaza, J., Gover, S. & Wolf, W. in *Molecular Replacement: Proceedings of the CCP4 Study Weekend* (ed. Dodson, E. J.) 87–90 (SERC, Daresbury, 1992).
- Cowan, K. in *Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography* **31**, 34–38 (1994).
- Nolte, R. T., Conlin, R. M., Harrison, S. C. & Brown, R. S. Differing roles for zinc fingers in DNA

recognition: structure of a six-finger transcription factor IIIA complex. *Proc. Natl Acad. Sci. USA* **95**, 2938–2943 (1998).

- Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
- Oberfield, J. L. *et al.* A peroxisome proliferator-activated receptor gamma ligand inhibits adipocyte differentiation. *Proc. Natl Acad. Sci. USA* **96**, 6102–6106 (1999).
- Ullman, E. F. *et al.* Luminescent oxygen channeling immunoassay: measurement of particle binding kinetics by chemiluminescence. *Proc. Natl Acad. Sci. USA* **91**, 5426–5430 (1994).

Acknowledgements

We thank B. Wisely and R. Bledsoe for making co-repressor expression constructs in early crystallization studies; W. Burkart and M. Moyer for protein sequencing; M. Iannone for compound characterizations; G. Waite and C. Wagner for mass spectroscopy and amino-acid content analysis; and J. Chrzas and A. Howard for assistance with data collections at 17-ID. Use of the Advanced Photon Source was supported by the US Department of Energy, Basic Energy Sciences, and Office of Science. We also thank L. Kuyper and D. Eggleston for support and critical reading of the manuscript.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.E.X.

(e-mail: ex11957@gsf.com). The Protein Data Bank code for the PPAR α /GW6471/SMRT complex and the PPAR α /GW409544/SRC-1 complex is 1KQQ and 1K7L, respectively.

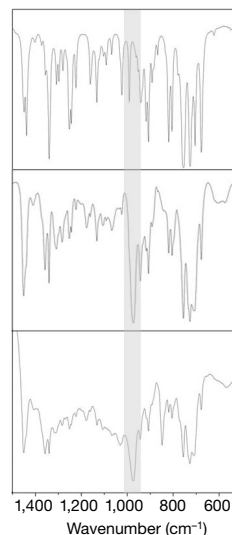
correction

Autonomic healing of polymer composites

S. R. White, N. R. Sottos, P. H. Geubelle, J. S. Moore, M. R. Kessler, S. R. Sriram, E. N. Brown & S. Viswanathan

Nature **409**, 794–797 (2001).

In this Letter, the middle infrared spectrum in Fig. 3b, corresponding to an authentic sample of poly(DCPD) prepared with Grubbs' catalyst and DCPD monomer, was a duplicate of the top spectrum owing to a formatting error. The corrected spectra are shown below. □



Turning the heat on PCR

Thermocycling and PCR are the themes this week.

Auto-Lid

Applied Biosystems

www.appliedbiosystems.com

Cover-up for System 9700 thermal cyclers

The Auto-Lid dual 384-well module for the GeneAMP PCR System 9700 thermal cycler is driven by a stepper motor that provides precise movement forwards and backwards. A plate ejection feature allows access by a robot arm and reduces the need for user intervention. The Auto-Lid can be purchased as an option with System 9700, or as a separate model to attach to an existing system. Applications include high-throughput PCR and cycle sequencing.



Auto-Lid, for high-throughput PCR.

RiboAmp

Arcturus

www.arctur.com

RNA amplification kit

The RiboAmp kit features a new linear amplification method designed to improve reliability. Reagents, nucleic acid purification devices and a protocol optimized to produce microgram quantities of RNA ready for labelling and hybridization onto microarrays are included with the kit. RiboAmp will amplify mRNA of all abundance classes, including low-abundance mRNA. Each kit contains supplies for 10 amplifications, and the amplified material can be utilized in multiple experiments. The kit can be used independently or as part of Arcturus's integrated systems for Microgenomics centred around its PixCell II Laser Capture Microdissection System.

Microfuge PCR rotor

Beckman Coulter

www.beckmancoulter.com

Capacity and power for DNA applications

The F12 × 8.2 PCR tube strip rotor for the Microfuge 22R microcentrifuge minimizes condensation from PCR reactions by providing quick spins and can be used for post-reaction clean-up. Up to 12 strips of eight 200 µl

tubes can be accommodated at a maximum of 20,800g. The rotor features an aerosol-tight lid and is autoclavable at 121 °C.

Amplification tape

Nalge Nunc

www.nalgenunc.com

Put a damper on it

Designed for NUNC 96- and 384-well amplification plates, this sealing tape can be used in temperatures ranging from -80 °C through to thermal cycling. It is certified RNase/DNase-free. Nalge Nunc claims that the tape provides a highly effective seal, preventing evaporation and eliminating well-to-well cross-contamination. The company's range of sealing tapes can be used with microwell plates in PCR, general assays, genomics, compound and library storage, bioanalytical assays, high-throughput screening and drug discovery applications.

MiracleHyb

Stratagene

www.stratagene.com

A clear message for northern analysis

MiracleHyb nucleic acid hybridization solution is said to yield high sensitivity and lower background, making it suitable for use in northern analysis when sensitive detection of low-abundance messages is a priority. The solution comes ready to use. It may be stored at room temperature and incorporated into existing protocols. A probe preparation buffer is included. For a limited time, Stratagene is offering a free 25 ml sample.

PROTEINscript-PRO

Ambion

www.ambion.com

High-yield translation kit

The PROTEINscript-PRO kit is designed for coupled *in vitro* transcription and translation using a highly active *E. coli* S30 extract. The format is a single-step reaction that can translate

both prokaryotic and eukaryotic sequences in 30 minutes. Any DNA template can be used, including PCR fragments. The prokaryotic lysate is said to produce higher yields than reticulocyte lysate for most sequences. The kits are available in several formats, including 25-reaction, single-tube kits using SP6, T7 or T3 polymerases and a 96-well high-throughput format using either SP6 or T7 polymerases.

SuperScript

Invitrogen

www.invitrogen.com

Single-tube RT-PCR system

This system combines SuperScript II reverse transcriptase with Platinum *Taq* DNA polymerase to provide RT-PCR in a single-tube reaction. Reduced RNase H activity in SuperScript II RT works with Platinum antibody-mediated, hot-start technology to provide high yield and sensitivity to 0.1 pg in gene expression analysis.

SuperSignal RPA III

Pierce

www.piercenet.com

Chemiluminescent detection kit

Pierce has collaborated with Ambion to produce this kit, which combines Pierce's SuperSignal chemiluminescent detection technology with Ambion's ribonuclease protection assay (RPA) to produce a non-isotopic RPA kit that provides sub-picogram sensitivity results in less time than it takes to run a traditional isotopic RPA. Exposures can be made in minutes and the probes remain stable for years, with no need to run probe purifications. The assay uses a single-tube reaction that requires no phenol extraction or proteinase K digestion.

HM-4000 Multidizer

UVP

www.uvp.com

Shake, rock, roll and rotate

The HM-4000 Multidizer hybridization oven features two independently operated chambers to accommodate hybridization and blotting procedures requiring different settings. The lower chamber is equipped with temperature controls, variable rotation speed settings and multiple-size bottle capacity with the use of a bottle carousel. The carousel can be removed for a rock-and-roll action using an optional rocker plate. The upper chamber has roller action for operating the shaker tray or an optional orbital tray. Applications include cDNA library screening, primer synthesis and nucleic acid hybridization.

These notes are compiled in the Nature office from information provided by the manufacturers.



Microfuge 22R: quick spins for PCR.

Contacts

Publisher: Fabien Savenay
Editor: Paul Smaglik
Sales Director: Ben Crowe

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Group European Manager:

Leonie Welss (4954)
European Manager:
Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953), Ben Corp (4974),
Andy Douglas (4975)
Holland/ Italy: Nevin Bayoumi (4978)
Scandinavia: Silke Opstrup (4994)
Spain/ Portugal: Leonie Welss
(4954)

Production Manager:

Billie Franklin
To send films and materials use
London address above.
Tel +44 (0) 20 7843 4814
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

International

Advertising Coordinator:

Laura Pearson (4977)

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Offices

France/ Belgium:

Christine Niox-Chateau
Tel + 33 (0) 1 43 20 16 51
Fax + 33 (0) 1 43 20 51 52
e-mail: c.nioxchateau@nature.com

Germany/ Austria/ Switzerland:

Patrick Phelan/ Kate Turner
Tel + 49 89 54 90 57 11/-2
Fax + 49 89 54 90 57 20
e-mails: p.phelan@nature.com
k.turner@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Director:

Ben Crowe
US Sales Manager: Peyton Mason

US Advertising Coordinator:

Ashly de Leon

Japan Head Office, Tokyo

MG Ichigaya Building (5F),
19-1 Harakakatamachi,
Shinjuku-ku,
Tokyo 162-0841
Tel +81 3 3267 8751
Fax +81 3 3267 8746
e-mail: k.cowan@naturejpn.com
Japan Manager: Kate Cowan

naturejobs

Putting pathogens first

If President George W. Bush's latest budget request is approved, basic immunology research in the United States will receive a shot in the arm. The plan, announced last week, would raise the budget at the US National Institute of Allergy and Infectious Diseases (NIAID) by 57% to \$4 billion, with the majority of this increase being directed at bioterrorism-related research. This is the single biggest non-defence item in Bush's bioterrorism proposal, which was prompted by the anthrax attacks that followed the 11 September hijackings and crashes.

Such an infusion of cash will create opportunities. But will they be constrained to bioterrorism applications? Anthony Fauci, NIAID director, offers assurances that the funds will pay for work "well beyond" the top five or six pathogens. "The entire field of infectious diseases and immunology will get a boost," he says.

The exact model for that boost remains unclear. But a contract between the NIAID and The Institute of Genomic Research (TIGR) in Rockville, Maryland, signed before the budget announcement, offers a hint of what is to come. The NIAID is paying TIGR to come up with tools to study pathogens whose genetic sequences are, or will soon be, known. This covers areas such as microarrays, genotyping, clone access and bioinformatics. TIGR will start with three pathogens and expand its scope as the project goes on.

It is safe to forecast that more pathogen sequencing and functional genomics will be in the pipeline, and that people who can contribute to those projects will be in demand. The unknown remains how individual investigators will be charged with using these data once they are produced. Will the resources be concentrated at a few centres or distributed more widely? With the amount of money available, it seems there would be room for both approaches — and the nature of the problem suggests that it would be prudent to support both.

Paul Smaglik

Naturejobs editor



Contents

REGIONS

Canada takes steps to
reverse its brain drain

p4

WWW.NATUREJOBS.COM

Career centre

Information on the scientific
job market

FOCUS

SPOTLIGHT

RECRUITMENT

SCIENTIFIC ANNOUNCEMENTS

SCIENTIFIC EVENTS



Saskatchewan's synchrotron (inset) is an example of the investment now being aimed at Canada's scientific infrastructure.

Drawing back the talent Canada

After postdoctoral stints at the University of Basel and the Massachusetts Institute of Technology, Michele Loewen was happy to come home to Canada last year. A full-time job in a government-funded lab, with access to a new national synchrotron was, she says, “an offer I couldn’t possibly say ‘no’ to”.

Loewen’s recruitment to Saskatoon’s Plant Biotechnology Institute illustrates the efforts of two institutions, the National Research Council (NRC), which runs the institute, and the Canada Foundation for Innovation (CFI), the government-created corporation that funds science-infrastructure improvements such as the University of Saskatchewan’s Canadian Light Source, a national synchrotron facility under construction in Saskatoon.

Although there is no central plan to reverse the brain drain — and no acknowledgment from Prime Minister Jean Chrétien that the problem exists (see *Nature* **401**, 731–732; 1999) — those institutions are showing signs that it is possible to retain Canadians who were thinking about leaving and to win back native scientists who have sought work elsewhere.

One reason that scientists, especially young ones, leave in the first place is because the pay does not match up — especially against top university posts in the United States and industry jobs in the United States, Europe and Japan. One CFI programme, the New Opportunities Fund, is addressing that imbalance by supplementing the salaries and lab budgets of newly hired faculty at Canadian universities. In four years, the programme has helped to hire more than 1,000 scholars — many of whom would probably have left otherwise.

One of them is John Valliant, a medicinal inorganic

chemist who has done promising work on new compounds for medical imaging and on using inorganic elements to treat diseases such as rheumatoid arthritis. After a fellowship at Harvard, Valliant was offered jobs at a US university and in the US pharmaceutical industry that would have paid him at least 50% more than Canadian offers. But the CFI programme allowed him to accept a position at

Back home: postdoc Michele Loewen returned to Canada after receiving an offer she couldn’t refuse.



High-tech stakes

Tony Patterson, founding editor of *Silicon Valley North*, a newspaper for Canadian high-tech communities, estimates that there are a million Canadians working abroad in technology-related fields. He calls them the country’s greatest untapped human resource and says that they “comprise a network from which skilled candidates might be recruited for re-integration to Canadian careers”.

To help this come about, he plans a “high-

tech homecoming” this year that will bring these “stars of Canada’s expat community” together with their national counterparts. He has enlisted the leaders of the industry to establish a website to help to market opportunities in Canadian high-technology to expatriates. The high-tech homecoming, he says, will “help define the vision to lead Canada in an age of rampant technological opportunity”.

◆ www.hightechhomecoming.com

D.S.

Attracting foreign nationals

Both public and private efforts are bringing foreign scientists into Canada.

On the public side, the National Research Council (NRC) recently recruited David Austin, a Briton who had spent nine years in Japan with the telephone company NTT. The NRC position appealed to him because it offered a better chance of a long-term research career than he was likely to have had in Japan.

On the private side, infusions of cash are also drawing some talent. Perhaps the most notable is the Perimeter Institute for Theoretical Physics, which was

established with Can\$120 million by high-tech entrepreneur Mike Lazaridis and two colleagues (see *Nature* 414, 391; 2001).

The new centre has brought Canadian Raymond Laflamme, an expert in quantum information and quantum computing, back from Los Alamos National Laboratories in the United States.

And it also attracted Fotini Markopoulou from the Albert Einstein Institute in Germany and Lee Smolin from Imperial College London and Pennsylvania State University.

D.S.

McMaster University in Hamilton, Ontario, without any financial qualms.

The Canada Research Chairs programme is taking the same approach to recruitment and retention, but at a more senior level. The government programme will have funded 2,000 positions in Canadian institutions by 2005.

Like the CFI programme, Canada Research Chairs has succeeded in luring back scientists who have spent some time working in other countries. After the programme's inception, several Canadian universities approached Vincenzo De Luca, a specialist in plant and tree biology who left the University of Montreal to join Novartis (later Syngenta Biotechnology) in the United States.

He eventually decided on Brock University, a small establishment in St Catherine's, Ontario, on the Niagara peninsula. De Luca's motivation for returning to Canada was his realization that he was really more interested in doing basic research on transgenic crops than product development. And that would be difficult in industry.

"The kinds of things I'm interested in are oriented towards pathways, to metabolic engineering," he says. "The other really important thing for me was my interest in trying to do something that would be of use to my country."

The Brock opportunity offered the chance to help solve regional problems connected with the grape

industry, which in Niagara generates millions of dollars and is important to tourism. De Luca thinks his industrial experience could be of use there.

Another important factor in De Luca's decision is that the chairs' appointments last for seven years, and are renewable for another seven after that. "You can imagine the amount of freedom that's giving me," he says. "And I feel that I'm at the right stage of my career to exercise a certain amount of leadership that can help a lot of the younger faculty."

Canada's largest research agency, the NRC, cannot benefit directly from either the CFI programmes or the Canada Research Chairs, but it is a major employer of scientists and has success in recruiting scientists from abroad (see left). This is partly because of its active participation in international scientific programmes. Richard Normandin, director general of the NRC's Institute for Microstructural Sciences (IMS), says the NRC is a sponsor of the Canadian-European Research Initiative on Nanostructures (CERION), which includes 17 European universities. About 30 NRC scientists have adjunct appointments outside Canada.

The NRC also benefited from a boost of Can\$40 million (US\$25 million) a year in the recent federal budget and Can\$110 million over three years for high-tech ventures such as its new National Institute of Nanotechnology (see *Nature* 412, 846; 2001). "We're in expansion mode," says Peter Hackett, the NRC's vice-president of research. "Hiring new researchers is now our number-one corporate priority."

One new NRC researcher was happy with the timing of his offer. The NRC offered Robert Wolkow a position at its microstructural institute just as the climate at Bell Labs in Murray Hills, New Jersey, where he then worked, was turning sour (see *Nature* 412, 578–579; 2001). Clinching the deal, the new position provided Wolkow with the opportunity to do pure science, just as Bell was pushing him to do mundane applied work.

For Wolkow and others who had worked outside Canada, it may have been the government programmes that helped to lure them back — but other issues have kept them happy.

Michele Loewen found living in Saskatoon less expensive than in Boston. She can afford things as a Canadian researcher that she couldn't as a US postdoc, such as her own house located 10 minutes' walk from the university, and frequent trips to the local concert hall. And there are other benefits. "The Rocky Mountains are not far," she says. "And we have clean air!"

David Spurgeon is *Nature's* Canada correspondent.

National Research Council

♦ hr.nrc.ca:8080/HRB/careersm.nsf/pagee/home

Industry Canada's SkillNet

♦ www.skillnet.ca

Biotechnology Human Resource Council

♦ www.bhrc.ca/About/index.html



Vincenzo De Luca came back to Canada under its Research Chairs programme.

Variety show: the chance to do more pure science convinced Robert Wolkow to return from the United States.



Recruiting: Richard Normandin.